PROCEEDINGS
OF THE
NATIONAL ACADEMY OF SCIENCES
Volume 20 January 15, 1934 Number 1

# *POLYISOMERISM AND ANISOMERISM IN CRANIAL AND DENTAL EVOLUTION AMONG VERTEBRATES*

By William K. Gregory

American Museum of Natural History, Department of Comparative and Human Anatomy

Read before the Academy, Tuesday, November 21, 1933

The researches of Beecher, Walcott and others on the Cambrian brachiopods show that some of these animals are extraordinarily generalized in appearance, in the sense that the body is composed of a large number of homodynamous segments, each equipped with similar appendages. Probably this condition is not as primitive as it looks and it would not be surprising if some day it may be proved that the remote ancestors of the trilobites had only a few segments, perhaps three. But for the purposes of the present paper the many segments of the trilobite *Triarthrus* may be assumed as the starting-point for the highly differentiated conditions illustrated in the typical crabs; in these some of the segments have fused to form a cephalo-thorax, while the paired appendages are greatly reduced in number but highly differentiated. We may thus take the conditions in *Triarthrus* as an example of what I shall here name *polyisomerism*, the state in which many homologous parts, or polyisomeres, are arranged along any primary or secondary axis, whether straight or curved. The highly differentiated condition in the typical crab, on the other hand, may be named anisomerism, that is, the state in which one or more parts are emphasized at the expense of the rest, while the original number of separate parts is usually reduced, either by fusion or by elimination.

Polyisomerism is a more inclusive term than *metamerism*, which has reference only to homologous parts arranged in an anteroposterior series, whereas in polyisomerism the axis may be anteroposterior, oblique, transverse or curved in any direction. Nor is it synonymous with the *meristic variation* of Bateson, which is concerned with numerical departures from the norm. Polyisomerism obviously results from a process analogous to or identical with budding, while anisomerism results from what Pézard and Huxley call *heterogony*, or unequal growth of one part of a polyisomerous series. Polyisomerism corresponds partly to what Cope in 1871[1] called anteroposterior repetitive acceleration.

The state or condition of polyisomerism may affect greater or lesser units of structure and it may come after or before anisomerism. A simple ascidian, for example, taken as a whole is highly anisomerous and extremely specialized for sessile life. A single *Salpa*, which is essentially a floating ascidian, is likewise largely anisomerous; but a *Salpa* chain is in a high degree polyisomerous and it has arisen through the budding of a stolon put forth by the parent *Salpa*.

It will be noted that this budding process tends to make the offspring extremely like its parent. In bisexual reproduction, on the other hand, the chromosomes, although polyisomerous in general appearance, unite to produce zygotes, which after passing through new cycles of poly- and anisomerism in development, give rise to individual differences within specific limits. From this viewpoint an ant colony, comprising several classes of related individuals all derived from one queen, is analogous with a complex organism of polyisomerous and anisomerous parts all derived from a single zygote.

Among vertebrates the processes under consideration affect all parts of the growing organism but are especially conspicuous in the integument and in the derm-bones underlying the integument. Usually the histological bases for polyisomerism are too small to be seen with the naked eye; but in the skull of certain highly specialized teleosts, such as *Aspicottus*, the processes of growth and differentiation have left an easily visible record. Here the centers of growth of the surface bones are revealed by the radiating rows of small bony papillae. The spikes that are characteristic of the cottids are evidently produced by the anisomerous emphasis of the posterior parts of certain otherwise polyisomerous growth centers.

Here we come to the principle that both polyisomerous and anisomerous parts always assume specific or individual characters, which are obviously determined in part by specific or individual hereditary factors. In certain fish skulls the forces that produce polyisomerism and anisomerism have left a record of the facts that they were extended in time, were repeated many times and were subject to recurrent accelerations and retardations. Here anisomerism is evidently due to the acceleration of certain parts of a polyisomerous series. Of course the real problem is, what determines the locus of each particular anisomerism. But while many factors doubtless enter, we cannot question that among these factors are the topographic relations of the part to the midline, together with its reactions to the surrounding parts. A comprehensive review of the history of fish skulls in geologic time[2] suggests that in addition to all the secondary factors of position, etc., each bone is the product of certain specific hereditary factors influencing its size, shape, texture and so forth.

Likewise in the field of the morphology of the skull of the tetrapodous vertebrates the concepts of polyisomerism and anisomerism are especially

FIGURE 1

**Contrast of relatively polyisomerous teleost skull with relatively anisomerous human skull.**

useful in describing the contrast between the generalized and the specialized. That the land-living vertebrates sprang from the base of the crossopterygian stem has been established by the researches of Traquair,

Smith Woodward, Watson, Bryant, Stensiö, Säve-Soderberg and others. Now the crossopterygians had a remarkably large number of skull bones, not less than one hundred and fifty, if we include the parts of the branchial apparatus. On the other hand, among the uppermost twigs of the vertebrate tree, in man, the number of separate skull bones in the fully adult skull does not normally exceed twenty-eight, as noted by the late Professor Williston. This reduction in number has occurred both by fusion and by elimination; the relatively few bones that remain are now markedly anisomerous (Fig. 1).

This progressive differentiation, or anisomerism, can be seen in all parts of the skull as we ascend the scale of vertebrates from fish to man, but it is especially conspicuous in the history of the upper and lower jaws. The mandible of crossopterygian fishes, as well as of the oldest tetrapods, includes no less than eighteen pieces; in adult man it appears as a single piece representing the coössified pair of dentaries of the lower vertebrates. The final dominance of the dentary in the lower jaw is matched by the progressive dominance of the maxillary in the upper jaw, which in the anthropoids and man annexes the premaxilla.

Regional anisomerism, when progressively developed phylogenetically, is measured by the changes in the proportional diameters of any given part. In the series of skulls from fish to man we observe in the later members the anteroposterior shortening and marked down-bending of the face, the vertical expansion of the braincase, the outgrowth of the nose and chin, etc.—all expressions of regional anisomerism.

It is well known that anthropologists have classified human crania according to the proportionate magnitude of one diameter in terms of another and that similar proportional differences of various parts of the head and body are denominated by terms indicating the relative height, width and length. In 1900 Professor Osborn used the terms *dolicho-*, *mesati-* and *brachy-cephalic* to describe the shape of the skull in the phyla of rhinoceroses, and later (1902) he applied the principle in distinguishing the numerous phyla of titanotheres. At first it was thought that dolichocephaly resulted from mesaticephaly by the uniform increase in length of all parts of the skull, but in 1904, as Professor Osborn's assistant, I was able to show that in several test cases the increase in length was not uniform throughout the skull but that the total lengthening took place by the summation of the different proportional increments of each part.[3]

Professor Osborn later invented the term *allometrons* to denote proportional measurements that change more or less independently in related phyla, as in the progressively brachycephalic or progressively dolichocephalic descendants of the primitive titanotheres. He also used the term *allometry* for that mode of evolution which is manifested in *allometrons*.

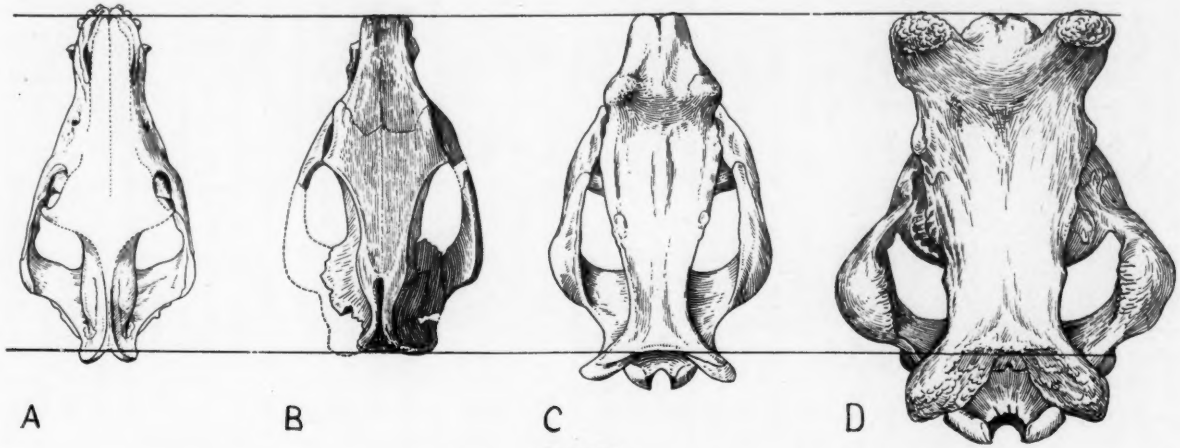


FIGURE 2

Transverse growth in titanothere skulls, culminating in extreme brachycephaly and localized outgrowths. Individual figures after Osborn but all reduced to same total length (tip of nasals to midoccipital crest).

A. *Eotitanops borealis*. Width index (top view) 51
B. *Manteoceras manteoceras*. Width index (top view) 58
C. *Brontops brachycephalus*. Width index (top view) 65
D. *Brontops robustus*. Width index (top view) 97

It was also believed that dolichocephaly could never give rise to anything but hyperdolichocephaly, but measurements of the skulls of fossil and recent perissodactyls prove that gigantic brachycephalic skulls have been derived eventually from small dolichocephalic skulls of the general type illustrated by known skulls of Lower Eocene perissodactyls (Fig. 2).

Similarly in the Proboscidea the extremely wide lower tusks of the *Platybelodon*, or shovel-tusked mastodonts, have evidently been derived from narrow tusks of the *Palaeomastodon* type. In brief, from the viewpoint of the present paper the phylogenetic principle named allometry by Professor Osborn is a special case of the more inclusive principle of regional anisomerism.

The study of the evolution of the teeth in vertebrates affords many examples of the ways in which a condition of primitive polyisomerism may give rise on the one hand to hyperpolyisomerism, or on the other to infinitely varied manifestations of regional anisomerism, involving changes of proportion, or allometry.

The presumably primitive condition of polyisomerism of the dentition may be seen in such Devonian sharks as *Cladoselache*, but in general the elasmobranchs run to hyperpolyisomerism.

A moderate degree of polyisomerism of the dentition is retained in the primitive reptiles. In the highly specialized *Ichthyosaurus*, on the other hand, the dentition has become hyperpolyisomerous. Among the serpents, there is a fair degree of polyisomerism in the pythons but in the highly specialized vipers the dentition is extremely anisomerous.

Among the mammal-like reptiles the cheek teeth early exhibited the phenomenon called cuspidation, which was carried to far greater development among the mammals. Here the growing tip of the dental germ has given off one or more buds, forming accessory cusps. When this tendency is unchecked it gives rise to various forms of multituberculism, or intradental polyisomerism.

In the early mammals the crown of the upper cheek teeth became extended transversely, but the number of main cusps was only three; later a fourth main cusp, the hypocone, was added. Among many lines of ungulates the crowns of the cheek teeth then lengthened anteroposteriorly. In certain lines of artiodactyls there was also a marked tendency for the sprouting of extra cusps and cuspules. In the family of the pigs the molar crowns lengthened rapidly anteroposteriorly, the main cusps tended to break up into cuspules and finally in the wart hogs the accessory cusps or conules greatly increased in number (Fig. 3, *F*, *G*).

The conules in the molar teeth of the more specialized Suidae are similar to those which appear (Fig. 3, *F'*, *G'*, *H'*) in certain phyla of Proboscidea and to which Professor Osborn has given the name of *aristogenes*. From the viewpoint of the present paper the proboscidean aristogenes are

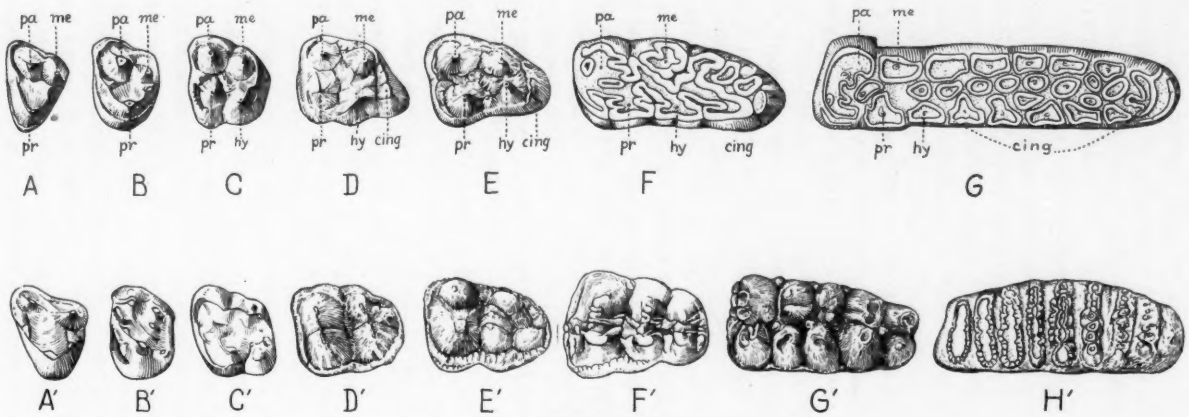


FIGURE 3

Third left upper molars, indicating progressive secondary polyisomerism, with lengthening of anteroposterior diameter and addition of ridges and conules in the later forms. All reduced to same width.

A. *Chriacus baldwini.* After Osborn.
B. *Protogonodon pentacus.* After Osborn and Earle.
C. *Palæochoerus typus.* After Stehlin.
D. *Palæochoerus waterhousi.* After Stehlin.
E. *Dicoryphochoerus* sp. After Colbert.
F. *Sus falconeri.* After Lydekker.
G. *Phachochoerus* sp.

A′. *Haplomylus speirianus.* After Matthew.
B′. *Hyopsodus powellianus.* After Matthew.
C′. *Phenacodus primævus.* After Matthew and Granger.
D′. *Moeritherium gracile.* After C. W. Andrews.
E′. *Palæomastodon minor.* After C. W. Andrews.
F′. *Serridentinus anguirivalis.* After Osborn.
G′. *Mastodon sivalensis.* After Falconer.
H′. *Stegodon bombifrons.* After Falconer.

polyisomeres, due to budding; this is also very evident in the upgrowth of new plates at the posterior end of the tooth and in the tendency for the main plates in the embryo elephant to break up into separate columns, like the conules of the wart hogs.

It will be seen that the multiplication of plates and conules lengthens the anteroposterior diameter of the tooth. In other words, in this case and many others allometry is associated with intradental polyisomerism.

*Conclusions.*—In conclusion, polyisomerism results from processes of cell division, budding or reduplication, anisomerism from heterogony, which implies regional or selective acceleration or retardation of polyisomerous units. The forces that produce polyisomeres are extended in time, repeated more or less frequently and subject to rhythmic acceleration and retardation.

The allometrons of Osborn, which imply phylogenetic changes in proportional diameters of any given organ or region, are an expression of anisomerism, or regional emphasis. The aristogenes of Osborn are polyisomeres, as in the added plates and conules of proboscidean molars; some of his rectigradations are polyisomeres, as in the case of new cusps, e.g., metastylids, added in similar positions in the cheek teeth of various perissodactyls; others, like the paired supraorbital swellings at the junction of the frontals and nasals in titanotheres, are anisomeres. The addition of conules and plates is bound to change the contour of a tooth and the outgrowth of horn swellings induces notable changes in the proportions of the skull roof. In other words, allometry is inextricably associated with rectigradation and both are special cases of polyisomerism and anisomerism.

Anisomerism results from unbalanced or disharmonic polyisomerism. When the dental lamina continues to bud off new teeth we call this condition polyisomerism. If, however, one tooth germ buds off an extra cusp, or if one cusp grows larger, it produces an effect of anisomerism or disharmony. A continuation of the process of budding off new ridges or cusps produces the condition here called intradental polyisomerism, as in the sectorial premolar of many multituberculates or in the many-plated molars of capybaras and proboscideans.

The morphological condition here termed polyisomerism occurs in all grades of organization from intra-atomic units to super-galaxies. Polyisomerism and its opposite derivatives, anisomerism and hyperpolyisomerism, evidently make possible the almost infinitely varied structural and physiological combinations that are recorded in the millions of varieties of plants and animals.

Adaptive radiation seems to have resulted from the summation along divergent lines of the results of secular polyisomerism, anisomerism and hyperpolyisomerism, while the production of allometrons, rectigradations and aristogenes appears to have been incidental to the same process.

Inspection of the paleontologic history of many phyla shows that when certain members of a series of polyisomeres have been subjected to increasing stresses through long periods of geologic time, they have often become enlarged, strengthened or fused with their neighbors and have thereby passed into a state of anisomerism.

Thus taxonomy, paleontology, comparative anatomy, genetics, experimental biology, physiology and allied sciences may find new common grounds in tracing the history and behavior of polyisomeres, anisomeres and hyperpolyisomeres.

[1] Cope, E. D., "The Method of Creation of Organic Forms," *Proc. Amer. Phil. Soc.*, 238, 241, 242 (Dec. 15, 1871).

[2] Gregory, William K., "Fish Skulls: A Study of the Evolution of Natural Mechanisms," *Trans. Amer. Phil. Soc.*, **23**, Art. 2 (1933).

[3] Osborn, H. F., "The Titanotheres of Ancient Wyoming, Dakota and Nebraska," *Monograph* **55**, *U. S. Geol. Surv.*, *I*, 830 (1929).

---

## *ACTIVE AND PASSIVE IMMUNIZATION IN TYPHUS FEVER*

BY HANS ZINSSER AND M. RUIZ CASTANEDA

DEPARTMENT OF BACTERIOLOGY AND IMMUNOLOGY
HARVARD UNIVERSITY MEDICAL SCHOOL

Read before the Academy, Tuesday, November 21, 1933

In preceding papers we have reported the successful production of a typhus vaccine made by obtaining concentrated suspensions of the Rickettsiae of the Mexican disease in the peritoneal cavities of rats in which the resistance has been reduced by radiation with x-rays. Since organisms of the Rickettsiae type cannot be cultivated, this method is so far the only one by which concentrated Rickettsiae suspensions can be obtained on a scale practically available for serological study and for extensive immunization. The Weigl method of producing similar suspensions for the same purpose by passing the European virus through lice and emulsifying the intestinal canals of the infected insects is excellent, but is hardly practicable on anything but a limited scale and is excessively dangerous, since the infected lice must be fed upon immunes for ten or more days after infection and must be handled under conditions of the greatest difficulty.

In the preceding papers alluded to we demonstrated that the vaccines now routinely prepared by us and by the Mexican Health authorities produce a satisfactory active immunization in guinea pigs infected with the Mexican virus and give about 30 to 50% immunization if the animals are sub-

sequently infected with the virus of the classical European disease. Serological work carried out by special methods of agglutination with these x-rayed rat vaccines indicated a close antigenic relationship between the two varieties of typhus virus mentioned, since human and animal convalescent sera from both types of the disease agglutinated both our Rickettsiae and the Weigl louse vaccines and since rabbits immunized with the two preparations, respectively, developed agglutinins for both Rickettsiae suspensions—the rabbit receiving our Mexican Rickettsiae powerfully agglutinating the Weigl vaccine, and vice versa. Additional proof of the specificity of our materials is furnished by the observation that the rabbits so treated developed Weil-Felix reactions.

In logical continuation of these results, we proceeded, two years ago, with the systematic treatment of a horse with our vaccines killed with 0.2% formaline, and we reported last year that the serum of this animal developed a Weil-Felix reaction and, although it had never received anything but the vaccine produced with the Mexican virus, it powerfully agglutinated the Weigl vaccines produced in lice with the European virus. At the same time, we showed that the serum of this horse, administered between twenty-four and seventy-two hours after infection with the virus, prevented the development of typhus in guinea pigs receiving heavy doses of the Mexican virus. These results are shown in a chart which has already been published[1] and which is important for the results which form the chief subject of our present report. The effect of this serum on typhus fever in man is being studied by a commission of Mexican physicians in their own country.

At the time that the last-mentioned experiments were done, we were unsuccessful in passively protecting guinea pigs with this serum against inoculations of the European virus. These results were anomalous and not easily explained in view of the close relationship between the two types of virus which the experiments outlined above had demonstrated. The most likely explanation seemed that our horse serum had not reached a sufficient potency to overcome the slight but definite antigenic difference between the two varieties of infectious agents. We continued, therefore, to treat the horse with increasing amounts of formalinized Mexican Rickettsiae, and resumed immunization experiments at the time when the horse's serum had attained an agglutinative potency of 1–640 against Bacillus Proteus X 19—about double the potency obtained in the bleedings used for the protection experiments just mentioned.

Also, we took into consideration in the experiments about to be presented the fact that when one inoculates European virus, the material is in the form of brain suspension and defibrinated blood of animals at the height of the disease. In such material the infectious agents are largely intracellular, and since the cells injected are species homologous, it is unlikely

that the virus enters the blood-stream of the inoculated animal in any considerable amounts until about four days after injection. Any serum given on the first day—as in previous experiments with the European material—is largely eliminated and is unavailable at the time when the infection seriously begins. We therefore altered our technique by allowing from two to three days to elapse between intraperitoneal inoculation of the virus and the subcutaneous administration of the horse serum. A preliminary experiment, in which four animals received a large intraperitoneal dose of European virus followed in two of them, after four days, by 2 cc. of the horse serum given subcutaneously and, on the fifth day, by an additional dose of 0.5 cc., gave us promising results in that the two controls developed typical and severe typhus fever and the untreated animals showed nothing more than temporary rises of temperature. Similar experiments are still going on.[1]

It is apparent from these simple tests that properly managed experiments in passive immunization—apart from the possible practical implications—corroborate our previous experience of the close immunological relationships existing between the classical European typhus fever and what is spoken of as the "New World" type.

Unlike many European observers, we are quite convinced that there is no fundamental difference between these two manifestations of Rickettsiae infection—that they belong essentially to the same group, with fractional antigenic differences probably brought about by passage through different animal and insect vectors.

That the European type of virus maintains itself for many years in America, largely in the immigrant population, was demonstrated by us recently by isolation from a case of Brill's disease occurring in Boston of a virus which has conformed through almost twenty guinea pig generations to the European type. Since the New World variety has been shown by ourselves and others to maintain itself in the interepidemic period in rats and rat fleas, and since the interepidemic reservoir of the classical European disease is not as yet known, there may be prospects by solving the reservoir of the European virus in American cities, of throwing light on the long mysterious question as to how the typhus virus maintains itself during long periods of quiescence in European communities where it breaks out so disastrously under conditions of famine, war and economic stress.

[1] The chart which illustrates the results of serum treatment in animals infected with Mexican virus was published in a paper by Zinsser and Castaneda in the *Journal of Experimental Medicine*, **57**, 395 (1933). The two charts illustrating the experiment here reported in a preliminary way will be published with the paper giving specific details of all the experimental work done on this subject.

# EVOLUTION OF THE EXPANDING UNIVERSE

BY G. LEMAITRE

UNIVERSITY OF LOUVAIN

Read before the Academy, Monday, November 20, 1933

The problem of the universe is essentially an application of the law of gravitation to a region of extremely low density. The mean density of matter up to a distance of some ten millions of light years from us is of the order of $10^{-30}$ gr./cm.$^3$; if all the atoms of the stars were equally distributed through space there would be about one atom per cubic yard, or the total energy would be that of an equilibrium radiation at the temperature of liquid hydrogen. The theory of relativity points out the possibility of a modification of the law of gravitation under such extreme conditions. It suggests that, when we identify gravitational mass and energy, we have to introduce a constant. Everything happens as though the energy *in vacuo* would be different from zero. In order that absolute motion, i.e., motion relative to vacuum, may not be detected, we must associate a pressure $p = -\rho c^2$ to the density of energy $\rho c^2$ of vacuum. This is essentially the meaning of the cosmical constant $\lambda$ which corresponds to a negative density of vacuum $\rho_0$ according to

$$\rho_0 = \frac{\lambda c^2}{4\pi G} \cong 10^{-27} \text{ gr./cm.}^3 \tag{1}$$

Let us consider the motion of matter symmetrically distributed round some fixed point 0. The classical equation of motion under the action of the modified gravitational field is

$$\left(\frac{dr}{dt}\right)^2 = -h + \frac{2Gm}{r} + \frac{\lambda c^2}{3} r^2 \tag{2}$$

where $m$ is the mass inside the sphere of radius $r$ and center 0. The condition that the system expands, remaining similar to itself, is that $h$ and $m$ have to be proportional, respectively, to $r^2$ and $r^3$. This classical motion is a good approximation of the relativistic equations when $r$ is small enough. When $r$ is great, some geometrical modifications become important and the classical model must be interpreted as a map in euclidean space. This map is like an orthogonal projection: lengths perpendicular to the radius vector are not altered, but along the radius vector they are represented at a scale

$$\sqrt{1 - h/c^2} \tag{3}$$

and the scale vanishes at the boundary of the map where $h = c^2$. If we write

$$h = c^2 \sin^2 \chi, \qquad r = R(t) \sin \chi, \tag{4}$$

we obtain Friedmann's equation, the variable radius $R(t)$ of the map being the so-called radius of space. Antipodal points at the boundary of the map are supposed to represent the same real points. Then it can be proved that if we change the center 0 of the representation the representation remains exactly the same.

The model we have described, we called the idealized model; it is perfectly homogeneous. If there are some fluctuations of density and velocity round the mean value, we can continue to apply classical mechanics (with the modified gravitational law) when we restrict ourselves to a domain not too large in respect to the whole volume of space. Let us suppose that the motion in the idealized model is of the ever-expanding type: a retarded expansion passing through a minimum velocity at some time $t_c$ when the total gravitation force vanishes, and expanding again under the predominant effect of the cosmical repulsion, the velocity tending finally to infinity according to the law

$$\frac{dr}{dt} = r\sqrt{\frac{\lambda c^2}{3}} = r\sqrt{\frac{4\pi}{3} G\rho_0}. \tag{5}$$

For the perturbed motion, i.e., for a distribution of mass and initial velocities somewhat different from the idealized model, the motion at some places may be of a completely different type from the motion of the idealized model. The relation between the energy-constant $h$ and the mass $m$ may be such that the motion is of the collapsing type: the expansion velocity vanishes when the gravitation is not yet completely balanced by the cosmical repulsion and the expansion is followed by a contraction. The result of the perturbations is that, after the time $t_c$, the system includes collapsing regions, distributed in the generally expanding space. That means that we obtain collapsing regions flying away one from another with velocities roughly proportional to the distance.

Occasionally, we may also have equilibrium-regions. The fact that such an equilibrium is unstable means only that it will occur relatively rarely and that collapsing regions will be decidedly more frequent than equilibrium-regions. Furthermore the equilibrium cannot be a detailed equilibrium, so that an equilibrium-region must divide itself into collapsing regions, and these collapsing regions will remain approximately at the same distance one from another.

The hypothesis we wish to discuss is that collapsing regions must be identified with the extra-galactic nebulae and the equilibrium-regions with the clusters of nebulae.

This hypothesis implies that the mean density in the clusters of nebulae must be the same for all, and furthermore must be connected with Hubble's

ratio of distance to spectroscopic velocities by the approximate relation (5). The epuilibrium-regions must have quite irregular forms, just as the clusters of nebulae have. For spherical form, we must have the relation

$$Nm = C D^3 d^3 . \tag{6}$$

where $N$ is the total number of nebulae in the cluster, $m$ the mean mass of a nebula for which we choose as a unit $10^9$ times the mass of the sun, $D$ the distance of the cluster in mega-parsecs, $d$ the angular diameter in degrees. The constant $C$ depends on the velocity $V$ in thousand km./sec. of the nebulae at a distance of a mega-parsec by the relation

$$C = \frac{3.084 \quad V^2}{8 \times 6.664 \times 1.983 \times (0.573)^3} = 0.155\ V^2. \tag{7}$$

In a previous paper,[1] I have compared the hypothesis of equilibrium with Hubble and Humason's data on eight clusters of nebulae.[2] In Hubble's work, the frequency distribution is determined and the distance deduced from the most frequent magnitude and checked by velocity determinations. This comparison is reproduced with some changes, explained later on, in table 1. Simultaneously with the publication of

TABLE 1

| CLUSTER | N | D | d | m |
|---|---|---|---|---|
| Coma | 800 | 19 | 1.7 | 1.1 |
| Perseus | 500 | 15 | 2.0 | 1.4 |
| Leo | 400 | 46 | 0.6 | 1.3 |
| Urs. Maj. | 300 | 30 | 0.7 | 0.8 |
| Cancer | 150 | 13 | 1.5 | 1.1 |
| Pegasus | 100 | 10 | 1 | 0.3 |

my paper quoted above, new data were published by Shapley[3] concerning 25 other groups of galaxies. The luminosity curve includes only the brightest members of the cluster and we can get an estimate of the total population and mean magnitude by identifying Shapley luminosity curves with the brightest part of the mean luminosity curve determined by Hubble. This mean luminosity curve is essentially an equilateral triangle of basis five magnitudes. We therefore try to represent Shapley luminosity curves by a straight line $a(m - b)$ and take as estimated population $169\ a/5$ and mean magnitude $b + 2.5$. This process gives very definite results for nine of the Shapley groups. They are tabulated in table 2. For nine other groups the process can be applied but with serious uncertainty as the fluctuations round the straight line are fairly large (table 3). For seven of the poorest groups the distribution has no appearance of similitude with Hubble's frequency curve and we have been obliged to neglect them. We have included under $n$ the observed number of nebulae

in Shapley's restricted area and taken for $d$ the angular diameter corresponding to his sin $\theta$. In the computation of both Hubble's and Shapley's data, we have used, following Shapley and Knox Shaw,[3] $-14.5$ as mean photographic magnitude of a nebula and 405 km./sec. for the spectroscopic velocity at a megaparsec.

Except for a systematic difference between Hubble's and Shapley's data, the constancy of the computed $m$ is very remarkable. This systematic difference corresponds to a change of 0.8 in the system of magnitude. This is what we may expect as Hubble and Shapley find, respectively, 13 and 14 as the most frequent magnitude of the Virgo cluster.

TABLE 2

| GROUP | n | N | D | d | m |
|---|---|---|---|---|---|
| 14 | 370 | 3800 | 46 | 1.57 | 2.0 |
| 20 | 341 | 1500 | 46 | 1.50 | 5.6 |
| 8 | 317 | 2700 | 44 | 1.67 | 3.7 |
| 15 | 256 | 3200 | 50 | 1.23 | 1.8 |
| 17 | 157 | 940 | 55 | 1.00 | 4.4 |
| 13 | 150 | 1300 | 46 | 1.17 | 2.9 |
| 3 | 77 | 270 | 44 | 0.67 | 2.4 |
| 25 | 56 | 340 | 63 | 0.53 | 2.8 |
| 4 | 46 | 340 | 50 | 0.67 | 2.7 |

TABLE 3

| GROUP | n | N | D | d | m |
|---|---|---|---|---|---|
| 16 | 90 | 340 | 44 | 0.87 | 4.1 |
| 2 | 88 | 170 | 30 | 1.00 | 4.0 |
| 23 | 82 | 300 | 44 | 1.00 | 7.1 |
| 21 | 66 | 240 | 44 | 0.60 | 1.9 |
| 10 | 58 | 170 | 36 | 0.83 | 4.0 |
| 22 | 58 | 300 | 44 | 0.50 | 8.8 |
| 5 | 56 | 90 | 27 | 0.50 | 0.7 |
| 7 | 42 | 150 | 33 | 0.50 | 7.7 |
| 11 | 45 | 150 | 45 | 0.60 | 3.3 |

We therefore find from the value of the red-shift and data on the clusters a mean mass of a nebula of one or three $10^9$ suns. This is the order of magnitude that was deduced by Hubble from the rotation and absolute magnitude of some bright nebulae.[4]

It might be noticed that our determination includes obscure matter. We can therefore conclude that, if our hypothesis can be accepted, the uncertainty factor of 100 or 1000 which is generally ascribed to Hubble's determination of the mean mass of a nebula must be considered as greatly exaggerated.

We have seen that our identification of the clusters of nebulae with equilibrium-regions is substantiated by observation. We must now inquire under what conditions we can identify the nebulae themselves

with the collapsing regions. A nebula of $10^9$ suns comes out of a collapsing region of initial volume corresponding to a radius $r_c$ given by

$$\frac{4\pi}{3}\rho_0 r_c^3 = m \tag{8}$$

that is, about 100,000 light years. For the radius of a normal nebula, we can take 1000 light years. This difference accounts for the degree of concentration observed in the nebulae. But what happens to the gravitational energy due to the contraction?

If, for definiteness, we consider the distribution of matter in a nebula as a polytrop of index 3, this energy is

$$\frac{3}{2} G \frac{N^2 m^2}{R} \tag{9}$$

where $R$ is the radius of the nebula and $N$ the number of stars of mass $m$. For comparison the gravitational energy of one star is

$$\frac{3}{2} G \frac{m^2}{r} \tag{10}$$

where $r$ is the radius of the star. Therefore the loss of energy is $Nr/R$ times the gravitational energy of the stars. For $N = 10^9$, $r = 6 \times 10^{10}$ cm. (radius of the sun) and $R = 1000$ light years $= 10^{21}$ cm. we find that we have to account for a loss of energy of the order of 6 per cent of the gravitational energy of the stars in the nebula. If the stars existed as stars before the critical instant when the collapse began, there would be no way to account for such a loss of energy. But, if before the critical instant matter is formed of gas, dust or meteorites of comparatively small free path, the collapse will produce a number of non-elastic collisions, turning out gravitational energy into heat, and agglomerating the diffuse matter into large hot masses, i.e., into stars.

We are therefore led to the conclusion that, in the frame of our hypothesis, the same mechanism which provides the formation of the nebulae, provides also the formation of the stars. Stars and nebulae are formed by the same process, and there is no star which is not associated with a nebula.

Another consequence is that obviously the total volume of the expanding and of the collapsing or equilibrium regions at the critical instant cannot be of different order of magnitude. It follows that the total mass of obscure matter in inter-nebular space is of the same order of magnitude as the total mass agglomerated into nebulae, i.e., $10^{-30}$ gr./cm.$^3$

The difference of types of the nebulae may be accounted for as a difference of the total angular momentum of the corresponding collapsing

regions. Finally, we must expect the occurrence of an intermediary type between the collapsing regions and the equilibrium regions. There must exist slowly collapsing regions containing a number of rapidly collapsing regions, and this could be identified with our galaxy.

We may expect to get a complete theory of all the problems connected with extra-galactic nebulae by applying statistical mechanics to small inhomogeneity in our idealized model. Such an investigation would probably involve only two parameters; one to fix the mean velocity of expansion at the instant of equilibrium, a second one to define the dispersion of the distribution of matter from the idealized model.

[1] *Ann. Société Scientifique Bruxelles*, Série **A 53**, p. 51–85 (1933). C. R. Paris, March 24 and April 10 (1933).

[2] Mt. Wilson Contribution No. 427.

[3] *Proc. Nat. Acad. Sci.*, **19**, 591–596 (1933).

[4] Mt. Wilson Contribution No. 324.

---

## *AN INVESTIGATION OF THE STRESSES IN LONGITUDINAL WELDS*[1]

BY WILLIAM HOVGAARD

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Read before the Academy, Wednesday, November 22, 1933

1. During the last four years a theoretical and experimental study of the stresses in longitudinal welds has been carried out at the Massachusetts Institute of Technology. The early part of this work has already been reported in these PROCEEDINGS, **16**, pp. 667–678 (1930) and **17**, pp. 351–359 (1931); the later part consists of three theses and a great amount of further study. The research had for its main object the determination of the stresses in a longitudinal weld as well as in the adjoining structural members, and centered in the simple and fundamental case where a long rectangular plate, subject to a lengthwise pull, is reenforced by a double flat bar or web connected to it by four fillet welds.

2. The problem was first attacked by the author, as described in the papers referred to above, by assuming that the shearing stress at any point in the weld is proportional to the average displacement of the bar relative to the plate across a transverse section through that point. Expressed in symbols:

$$q_x = U_x/\mu \qquad (1)$$

where $q_x$ is the shearing stress on the throat area of the weld. $U_x$ is the relative displacement and $\mu$ is a coefficient, hereafter referred to as the

"displacement coefficient," which was assumed to be constant all along the weld in a given structure.

The following formula for the shearing stress was then obtained by applying the Principle of Least Work and the Method of Variation.

$$q_x = \frac{p \sinh mx}{m\mu E \cosh mL} \tag{2}$$

where

$$m = \sqrt{\frac{\gamma(A + a)}{Aa\mu E}} \tag{3}$$

$p$ = the terminal stress in the plate due to the pull of the testing machine
$L$ = one-half length of bar or web
$A$ = sectional area of plate
$a$ = sectional area of bar or web
$\gamma$ = throat area of weld per unit length
$E$ = modulus of elasticity, assumed to be the same for the entire structure including the weld.

Later a simpler proof of the same formula was developed by Professor H. E. Rossell.

To check formula (2), the author with the assistance of Professor I. H. Cowdrey and Professor R. G. Adams carried out a great number of tests, involving more than two thousand strain measurements on a plate reenforced by flat bars. A preliminary value of the displacement coefficient was determined and valuable experience was gained as to the form and preparation of the test specimen. The difficulties connected with these experiments showed the necessity for adopting a test piece provided with normal webs and for making a more complete theoretical study, especially of the state of stress in the plate and in the web.

3. *The first attempt* in this direction was made by three Naval Students: R. D. Conrad, R. A. Hinners and L. V. Honsinger, all Lieutenants (j. g.) in the Corps of Naval Constructors, who, under the supervision of Dr. H. Hencky, prepared a thesis on this subject.[2] In order to simplify the problem the web was not connected to the plate by continuous lines of welds, but only by spot welds at each end and one at the middle for positioning. Thus approximately two single forces came to act in opposite directions, one at each end of the web, creating two nuclei of strain. The stresses caused by a single force operative at a point in a plate of infinite extent was already known,[3] and an Airy stress function $F_I$ was found by integration. This function must satisfy the biharmonic differential equation:

$$\Delta^4 F_I = 0 \tag{4}$$

and should be so modified as to satisfy also the boundary conditions.

This was effected by adding various functions $F_1$, $F_2$, $F_3$, etc., to it, each of which must at the same time satisfy the biharmonic and the boundary conditions.

Apart from the pull of the testing machine which was dealt with separately, the boundary conditions were as follows:

$$\left.\begin{aligned} &\text{At the ends where } x = 2L, \text{ the longitudinal stress } \sigma_x = \frac{\partial^2 F}{\partial^2 y^2} = 0 \\ &\text{At the sides where } y = \pm b \text{ the transverse stress } \sigma_y = \frac{\partial^2 F}{\partial x^2} = 0 \\ &\text{For all boundaries the shearing stress } \tau_{xy} = \frac{-\partial^2 F}{\partial x \partial y} = 0 \end{aligned}\right\} \quad (5)$$

It was easy to find functions which conformed to these conditions, but it proved impossible at the same time to satisfy the biharmonic equation, and therefore another solution was worked out by which the boundary conditions were only approximately fulfilled, while the biharmonic was satisfied rigorously.

It was found that three-eighths of the total load was transmitted through the web in this case. A limited number of strain measurements subsequently made showed a fair correspondence with the theoretical results.

4. *The next step in this investigation* was made by a student in the Course of Naval Architecture and Naval Construction, Dr. Y. C. Yeh, who made a very comprehensive study of the problem in preparation of a thesis for the degree of Doctor of Science.[4] In this case the welds were extended the whole length of the web, and the state of stress was studied for the web as well as for the plate. The first part of the thesis was prepared under the guidance of Dr. Hencky.

The weld was assumed to be subject only to shearing and was regarded as an internal boundary of the plate as well as of the web, common to both. It was found convenient to employ elliptic coördinates:

$$\left.\begin{aligned} x &= L \cosh \alpha \cos \beta \\ y &= L \sinh \alpha \sin \beta \end{aligned}\right\} \quad (6)$$

where the curves $\alpha$ = constant are confocal ellipses and $\beta$ = constant represent a family of hyperbolas. The ellipse $\alpha = 0$ is identical with the weld line.

The general expressions for stresses and displacements are given by Professor C. E. Inglis in a paper entitled "Stress in a Plate Due to the Presence of Cracks and Sharp Corners,"[5] and is also worked out by the use of stress functions in Coker and Filon's *Treatise on Photoelasticity*.

The external boundary conditions for large or infinite $\alpha$ are that the plate is subject to a uniform pull in the $x$-direction, while the web is free

from external forces. The internal boundary conditions at $\alpha = 0$ are that the displacements and the shearing stresses in each of the connected members and in the weld shall be the same at any point. Moreover, the transverse displacements shall be zero. In order to satisfy these conditions it was necessary to introduce five unknown coefficients in the plate solution and three in the web solution.

For the shearing stress along the weld line the following simple expression was obtained:

$$q_x = \frac{4}{7}\frac{t}{\gamma}p\frac{x}{\sqrt{L^2 - x^2}} \tag{7}$$

the stress being reckoned on the throat area per unit length of the weld.

In this formula the experimental values $E = 30.2 \times 10^6$ lb. per sq. in., and Poisson's ratio $1/(3.64)$ were used.

The displacement in the weld was found:

$$(u)_{\alpha=0} = -0.4760\, px/E \tag{8}$$

It is seen that according to this theory the shearing stress is infinite at the ends of the weld, but the displacement is finite.

It remained to correct the state of stress for the presence of the finite external boundaries. First the stresses existing at these bounderies in the infinite plate were calculated, hereafter referred to as the residual stresses, and then compensating stress functions were constructed which should annul these stresses and which at the same time should satisfy the biharmonic equation. It was found very difficult, however, to correct for the residual stresses on account of the interaction between the stresses at the external boundaries and the shearing stress in the weld line, and hence another mode of attack was adopted.

It was assumed that the expression for the shearing stress in the weld, equation 7, obtained for the infinite plate and web, could be used, provided the coefficient 4/7 was modified so as to conform to the experimental results. Now, it was found experimentally that the load was equally distributed between the plate and the web so that the expression for the shearing stress became:

$$q_x = \frac{1}{4}\frac{tp}{\gamma}\frac{x}{\sqrt{L^2 - x^2}} \tag{9}$$

In this way the interaction between the stresses on the external and internal boundaries was discounted, and in the correction for the residual stresses only the external boundaries would need to be considered.

The stresses and the displacements for the infinite plate and web were determined by the method of complex integration.[6] First the following expressions for the displacements were obtained:

$$\left.\begin{aligned} u &= -\frac{m+1}{4m}\,y\psi + \frac{\Phi}{2} - \frac{m+1}{4m}\,\frac{\partial U}{\partial x} \\ v &= -\frac{m+1}{4m}\,y\varphi + \frac{m-1}{4m}\,\Psi - \frac{m+1}{4m}\,\frac{\partial U}{\partial y} \end{aligned}\right\} \qquad (10)$$

from which the stresses and then the stress function

$$F_{II} = \frac{E}{4}(y\Psi + U)$$

was found.

Here $\varphi$, $\psi$, $\Phi$, $\Psi$ and $U$ are harmonic functions which are to be determined on the basis of the known shearing forces acting in the weld. For any pair of elemental forces, acting on a length $dx$ at distances $\pm x$ from the middle of the weld, a complex stress function $df(z)$ can be written down[7] and then the complete function $f(z) = \varphi + i\psi$ as well as $\varphi$ and $\psi$ can be found by integration. Another integration gives $\Phi$ and $\Psi$ and, finally, the potential function $U$ is determined so as to make the solution unique. All the functions were expressed in elliptic coördinates, and by substitution in (10) the displacements and then the stresses were determined for the infinite plate.

The correction for the residual stresses existing at the boundaries $y = L/4$; $x = 2L$ for the plate and $x = L$ for the web was now obtained by means of various compensating stress functions, which, when superposed on $F_{II}$, at least approximately fulfilled the actual boundary conditions.

In case of the plate two compensating functions were required, so that the solution became:

$$F = F_{II} + X_1 + X_2. \qquad (11)$$

In case of the web, three functions were needed and the work was more laborious and difficult.[8] The result was

$$F' = F_{II}' + X_1' + X_2' + X_3' \qquad (12)$$

where $F_{II}' = -(F_{II} + X_1)$.

For a complete solution, several other stress functions are required, and were developed, but as the respective corrections proved insignificant, they are not here included.

Now the stresses could be finally determined and it became possible to find the value of the displacement coefficient $\mu$, which was plotted in a curve and its average value $0.158 \times 10^{-6}$ was obtained.

Dr. Yeh studied also the effect of tapering the web at the ends by considering it to be bounded by an ellipse, assuming for simplicity that the plate was of infinite extent.

This made the shearing stress nearly 50 per cent smaller than for the rectangular web, but the terminal stress was still infinite. It is the intention to study this case experimentally in the near future.

5. *The third thesis*[9] was an extensive experimental study of the stresses in the same test specimen, made by Lieutenant A. M. Zollars (CC), U. S. N. of the Course of Naval Construction, and Mr. E. P. Worthen of the Course of Naval Architecture. The object was to obtain a picture of the stress field in both plate and web and to make a critical comparison with the theoretical results given above. Strains were measured with Huggenberger tensometers at a great number of points, a rosette of four measurements being taken at each point, from which the stresses were computed.

The longitudinal average stresses $\sigma_{x\,\text{avg.}}$ across transverse sections of the plate and web were determined graphically, whereupon the displacement $U_x$ and the shearing stress $q_x$ could be found for any point. Now the experimental value of the displacement coefficient $\mu$ was determined and its average was found to be $0.224 \times 10^{-6}$, as against $0.158 \times 10^{-6}$ by Dr. Yeh's theory. The curve representing Dr. Yeh's formula (9) falls considerably below the experimental values in the important outer part of the weld, until very near the end of the weld where it rises to infinity.

6. Actually the stress cannot rise above the yield point and it is proposed therefore to modify (9) in such a way as to make the curve terminate at the yield point, while elsewhere the ordinates are so increased as to maintain the area subtended by the curve unaltered. This must be so because this area is proportional to the total pull transmitted by the weld. The following formula is proposed:

$$q_x = \frac{ptr}{4\gamma} \frac{x}{\sqrt{k^2L^2 - x^2}} \tag{13}$$

where $k$ and $r$ are coefficients to be determined experimentally or by experience. In the present case it is found that $k = 1.006$, $r = 1.115$.

The question of shearing stresses in welds is one of increasing practical importance. The solution here proposed is not offered as a finality, but rather as the first step in a new line of research.

[1] A more detailed account will appear in the *M. I. T. J. Math. & Phys.*, **13**, No. 2 (1934).

[2] "Stress Field of a Plane Plate Reënforced by a Longitudinal Girder and Subjected to Tension," *M. I. T.* (1932).

[3] Love, *Th. of Elast.*, 3 ed., p. 207.

[4] The Distribution of Stresses in Welded Structures," *M. I. T.* (1933).

[5] *Inst. Nav. Arch.*, I, pp. 219–230 (1913).

[6] L. Föppl, "Konforme Abbildung ebener Spannungszustände," *Zeits. angew. Math. Mech.*, 81–92 (1931).

[7] A. and L. Föppl, *Drang und Zwang*, I, p. 273 et seq. (1920).

[8] In the correction for the stresses at the end boundaries, Dr. Timoshenko's Approximate Strain-Energy method was used.

[9] "An Experimental Determination of the Distribution of Longitudinal Shearing Stresses in a Continuous Weld," *M. I. T.* (1933).

---

## *ON THE DYNAMICAL THEORY OF ELECTRICAL COMMUTATOR MACHINES*

BY W. H. INGRAM

CITY COLLEGE, NEW YORK

Communicated December 4, 1933

The analogies between a system of particles in dynamics and a nexus of electric circuits, established by Maxwell, have not been pursued successfully hitherto to the case of commutator electrical machinery although the case of slip-ring electrical machinery has been rather exhaustively treated. It is the purpose of this note to point out that the relative angular velocity of commutator and brush plays a rôle in the theory of a commutated nexus corresponding to that played by the angular velocity of a system of moving coördinates in dynamics.

The equations of motion of a slip-ring machine, given by the Lagrangean formula,[1] may be written in the form

$$E_i = R_{ij}\dot{q}^j + L_{ij}(\theta)\ddot{q}^j + \Gamma_{i,\,mn}\dot{q}^m\,\dot{q}^n \qquad (1)$$

where $\Gamma_{i,\,mn}$ is the ordinary Christoffel symbol and where the coördinates are true. Certain of the $\dot{q}$'s represent the armature currents; for example, the first three in the case of a three-phase alternating-current motor. In the theory of this machine, a transformation to quasi-coördinates has been used[2] having the form

$$\left.\begin{aligned} \dot{q}^i &= \beta^i_k\,(\theta)\,\dot{\xi}^k \qquad (i = 1, 2, \ldots n-1) \\ \dot{q}^n &= \dot{\xi}^n = \theta \end{aligned}\right\} \qquad (2)$$

where $\theta$ is the privileged rotor position coördinate and where

$$\begin{aligned} \beta^k_1 &= {}^2/_3 \sin(\theta - 2\pi k/3), \qquad (k = 1, 2, 3) \\ \beta^k_2 &= {}^2/_3 \cos(\theta - 2\pi k/3), \qquad (k = 1, 2, 3) \\ \beta^k_3 &= {}^1/_3, \qquad (k = 1, 2, 3) \end{aligned}$$

$$\beta^k_j = \begin{cases} 1, & k = j = 4, 5, \ldots n-1 \\ 0, & k \neq j = 4, 5, \ldots n-1 \end{cases}$$

In terms of the new coördinates, the equations of motion may be written

$$e_i = r_{ij}\,\dot{\xi}^j + l_{ij}\,\ddot{\xi}^j + \lambda_{i,\,jk}\,\dot{\xi}^j\dot{\xi}^k + L_{ij}\,\frac{\partial\beta_k^j}{\partial\theta}\,\dot{\xi}^k\dot{\theta} \qquad (3)$$

where $\lambda_{i,\,jk} = \Gamma_{i,\,mn}\beta_j^m\beta_k^n$. For an orthocyclic machine, i.e., for a machine in which

$$L_{ij} = l^*\delta_{ij} - l_m - l_\mu\cos(2\theta - (i+j)2\pi/3), \quad (i, j = 1, 2, 3)$$

where $\delta_{ij}$ is the ordinary Kronecker symbol, it is found that all the $\lambda_{i,\,jk}$'s are zero for all $i$ except $i = n$.

The system (3), for these $\beta$'s, is known to be identical with that for a commutator machine with an infinite number of segments.[3] It is now found to be identical with the equations of motion of a Lagrangean system whose kinetic energy is expressed in terms of the $\dot{\xi}$'s and $\theta$, when the $\xi$'s are regarded as moving coördinates. Thus, given the functions of specification (energy, dissipation and activity[4]) in the true variables, in terms of the new coördinates they become

$$\begin{aligned} T &= {}^1/_2\, l_{\alpha\beta}\dot{\xi}^\alpha\dot{\xi}^\beta \\ S &= {}^1/_2\, r_{\alpha\beta}\dot{\xi}^\alpha\dot{\xi}^\beta \\ U &= E_i\beta_\alpha^i\dot{\xi}^\alpha = e_\alpha\dot{\xi}^\alpha \end{aligned}$$

and it is readily found that the equations (3) may be obtained from a generalization of the Lagrangean formula to systems with moving axes[5] on the assumption that the $\xi$'s form a system of coördinates rotating with the angular velocity $\dot{\theta}$. This establishes the theorem.

[1] Cf. Ingram, *Jour. Franklin Inst.*, **210**, 330 (1930).

[2] Park, *Trans. Am. Inst. Elect. Engrs.*, **48**, 716 (1929).

[3] Pointed out by G. Kron, "Tensor Analysis of Rotating Machinery" (1933), privately printed; discussed in detail by the writer in a paper with the title "Note on the Dynamical Theory of Commutated Motors" in a forthcoming number of the *Phil. Mag.*

[4] Ingram, *Ibid.*

[5] Cf. Whittaker, *Analytical Dynamics*, 3rd Ed., p. 216.

# THE NATURE OF THE DEPRESSOR SUBSTANCE OF THE BLOOD

BY CYRUS H. FISKE

BIOCHEMICAL LABORATORY, HARVARD MEDICAL SCHOOL

Read before the Academy, Tuesday, November 21, 1933

Zipf[1] has recently shown that fresh defibrinated blood and adenylic acid isolated from it have identical physiological properties, to the extent that both lower the blood pressure (largely as a result of vasodilation), stimulate the uterus and inhibit intestinal contractions. The effect on blood pressure is obtained also in animals under the influence of atropine and in rabbits anesthetized with ether, properties which (together with inhibition of the intestine) serve to distinguish adenosine derivatives from acetyl choline and histamine.

Since the adenylic acid first isolated from muscle in 1927 by Embden and Zimmermann[2] has proved to be a decomposition product of adenosine triphosphate, and not a natural constituent of this tissue, the question arises whether the same may not be true of blood. A portion of the organic acid-soluble phosphorus of blood has long been known to be readily hydrolyzed by acid,[3] as is the case with two of the three phosphoric acid groups of adenosine triphosphate, but the only previous attempt to isolate this substance from blood[4] has given no conclusive result.

In a preliminary note, Zipf[5] speaks of having secured experimental evidence in support of his belief that the depressor substance of the blood comes into play in the vasodilatory mechanism operating in hyperemia caused by heat, inflammation, etc. This can be true only if some means is provided whereby the active substance may be liberated from the corpuscles, where it is normally confined, under the influence of conditions which give rise to vasodilation. The complete elucidation of the chemical aspects of this phenomenon—assuming Zipf's contention to be correct—can hardly be expected in the absence of precise information regarding the nature of the depressor substance as it exists in the circulating blood.

The method previously used in this laboratory[6] for the separation of adenosine triphosphate from protein-free muscle filtrates consists in the successive precipitation of the compound with (1) mercuric acetate from 2 per cent acetic acid solution, and (2) calcium chloride and 4 volumes of alcohol. (For analysis the resulting acid calcium salt was converted to the acid silver salt by precipitation with silver nitrate from approximately 0.06 *N* nitric acid solution.) This procedure cannot be followed in the presence of any considerable amount of diphosphoglyceric acid, which is the principal organic phosphorus compound of the blood, for it

also is precipitated by both mercuric acetate and alcoholic calcium chloride. The separation from the glyceric acid ester can, however, be effected by means of the above reagents (mercuric acetate, silver nitrate and alcoholic calcium chloride) used in a different way. A brief outline of the method follows.

The blood is first treated with 4 volumes of ice-cold 10 per cent trichloroacetic acid to remove the protein. The protein precipitate is filtered off at 0°, and the filtrate promptly neutralized with a saturated solution of sodium hydroxide. One-twentieth volume of 20 per cent mercuric acetate is added to the neutralized filtrate, and the precipitate is separated by centrifugation, washed with water and decomposed with hydrogen sulphide. Since hydrolysis occurring during the process of isolation not only diminishes the yield but also affects the composition of the final product, the entire operation is best conducted at a low temperature and with the least possible delay. The risk of decomposition during the manipulation of the mercury precipitate is comparatively slight, but in the steps to follow the material must unconditionally be maintained at a temperature of about 0° throughout the process.

The cold filtrate and washings from the mercuric sulphide precipitate are treated with one-tenth volume of ice-cold 5 *N* silver nitrate solution. The precipitate, centrifuged after half an hour and washed with ice-water as long as the washings remain clear, is decomposed with hydrogen sulphide. The silver precipitation is now repeated in the same manner, but this time the silver is removed in the presence of a limited amount of water, so that the filtrate and washings from the silver sulphide precipitate occupy a volume only about one-eighth that of the original blood.

For the final step, the acid calcium salt is precipitated by adding one-tenth volume of 20 per cent calcium chloride ($CaCl_2.2H_2O$) and 5 volumes of ice-cold absolute alcohol. On the following day the precipitate is filtered off with suction, washed with absolute alcohol and dried in vacuo as rapidly as possible.

An estimate of the amount of adenosine triphosphate actually present in the blood can be made by taking advantage of the fact that nucleotides are completely adsorbed from dilute aqueous solution by charcoal (norit). Charcoal also adsorbs nucleosides, as well as the free purine and pyrimidine bases,[7] but apparently no other organic phosphorus compound is adsorbed from blood filtrates unless the norit is added in excess. For example, the amount of total organic phosphorus adsorbed from the neutralized protein-free filtrate prepared from one sample of rabbit blood was 6.79 mg. (per 100 cc. of blood), and of this 4.55 mg. could be hydrolyzed by heating for 15 minutes with *N* hydrochloric acid at 100°. The ratio of "hydrolyzable" to total organic phosphorus in the adsorbed material was consequently 67.0 per cent, a figure agreeing within the experimental

error with the 66.7 per cent to be expected from the fact that two-thirds of the phosphorus in adenosine triphosphate is set free as *o*-phosphoric acid by this treatment. The yield of acid calcium salt from this sample of blood was 6.45 mg. per cent (in terms of phosphorus), a recovery of 95 per cent.

In its behavior with the various reagents mentioned, the product isolated from fresh rabbit blood is indistinguishable from adenosine triphosphate prepared from muscle. The composition of the acid silver salt from both sources is likewise identical ($C_{10}H_{13}N_5O_{13}P_3Ag_3$). More extensive comparisons have not been possible up to the present time, owing to the fact that the intact substance has been obtainable only by using blood drawn directly from the animal into ice-cold trichloroacetic acid. In the product isolated from rabbit blood which has been collected in this way, the atomic ratio of nitrogen to phosphorus is 5 to 3, and all the nitrogen (within the error of analysis) is in the form of adenine.

Isolation experiments conducted on a comparatively large scale with slaughter-house material do not give the same result. Thus, in one case, 0.5 gm. of acid calcium salt was obtained from a sample of protein-free filtrate representing 1 liter of pig blood, but although the blood was transported from the slaughter-house as rapidly as possible in cold weather, the nucleotide was partly hydrolyzed by the time it had reached the laboratory, and in addition 8 per cent of the adenine present in nucleotide combination had been deaminized to hypoxanthine. The corresponding hypoxanthine derivative cannot be separated from unaltered adenosine triphosphate by any method known at present.

[1] Zipf, K., *Arch. exp. Path. u. Pharmakol.*, **160,** 579–598 (1931).

[2] Embden, G., and Zimmermann, M., *Z. physiol. Chem.*, **167,** 137–140 (1927).

[3] Zucker, T. F., and Gutman, M., *Proc. Soc. Exp. Biol. and Med.*, **20,** 133–136, 372–375 (1922–23).

[4] Barrenscheen, H. K., and Filz., W., *Biochem. Z.*, **250,** 281–304 (1932).

[5] Zipf, K., *Arch. exp. Path. u. Pharmakol.*, **167,** 60–61 (1932).

[6] Fiske, C. H., and Subbarow, Y., *Science*, **70,** 381–382 (1929).

[7] The following substances have been tested (in approximately 0.001 *M* solution, shaking for 5 minutes with 0.3 gm. of acid-washed norit per 10 cc.): adenosine triphosphate, adenylic acid, guanylic acid, cytidylic acid, uridylic acid, adenosine, adenine, guanine, uracil and cytosine. With the particular sample of charcoal used, 96 per cent of the uracil and 93 per cent of the cytosine were adsorbed; adsorption of the other compounds was at least 99 per cent complete.

# EFFECT OF X-RAYS ON THE RATE OF CHANGE IN THE UNSTABLE MINIATURE-3 GENE OF DROSOPHILA VIRILIS

BY M. DEMEREC

DEPARTMENT OF GENETICS, CARNEGIE INSTITUTION OF WASHINGTON, COLD SPRING HARBOR, N. Y.

Communicated December 5, 1933

It is well known that x-rays increase the rate of genetic changes, both of the chromosomal and of the genic ones. The experimental evidence indicates that the effect of x-rays is proportional to the dosage used. Experiments made with treated males and using sex-linked lethals as a criterion showed that a 1000 *r*-unit exposure increased the rate of change about ten times (Demerec, 1933*a*).

Several genes are known in *D. virilis* to be normally unstable, changing to their wild-type allels at a high rate. It was of considerable interest, therefore, to determine whether or not the rate of change of these genes would be affected by the x-ray treatment in a similar way as the rate of change in the so-called stable genes. The results of an experiment dealing with that problem are reported in this paper.

*Material.*—The gamma line of the unstable miniature-3 was used in these experiments. The flies were descendants of the material used in testing the effect of temperature on the rate of change in the miniature gene (Demerec, 1932). However, during the time which elapsed between the two experiments, a change from the gamma form to the alpha form had occurred so that the material was not a pure gamma line but a mixture of gamma and alpha forms.

It has been shown previously that gamma is unstable in somatic cells only while alpha is unstable in both somatic and germinal tissues. In addition alpha has a slightly higher rate of somatic change than gamma, which fact is responsible for the difference in the number of mosaics in the controls in the reports of the two experiments.

*Methods.*—Ten sets of flies with ten pairs in each set were used as parents. They were allowed to lay eggs in culture bottles during a 24-hour period after which time they were transferred into fresh culture bottles. Altogether fifteen transfers were made for each set of parents. Some of these transfers were used as controls and others were used for x-ray treatments.

The flies were raised in small glass jars, 4 centimeters in diameter and 5.5 centimeters high which were filled with corn-meal-agar food to the height of about two centimeters from the bottom.

The x-ray irradiation was given to larvae at different stages of ontogeny. Some sets were treated 24 hours after the parent flies were placed in the

culture bottles, viz., immediately after the parent flies were transferred to other cultures (egg stage); other sets were treated four days after the parents were placed in cultures (larval stage) and still others were treated seven days after that date (late larval and early pupal stage). The dosage was measured with a Fricke-Glaser dosimeter by keeping during the exposure a 1/20 cubic centimeter chamber adjacent to the treated cultures. The dosage applied was about 1800 *r*-units. Care was taken to have as large an area of the food as possible directly exposed to x-rays. The measurement indicates that the glass of the containers used absorbed about two-thirds of the rays and therefore the larvae and the pupae which happened to be shaded during the exposure received only about 600 *r*-units.

*Results.*—From table 1, which summarizes the results of this experiment, it may be seen that the x-ray treatment did not affect the unstable miniature-3 gene in the same degree as it affects other genes. The difference between the percentage of mosaics among total treated flies and controls is only 2.36 ± 1.09.

TABLE 1

NUMBER OF MOSAICS AMONG MINIATURE-3 FLIES TREATED WITH 600 TO 1800 *r*-UNITS AT DIFFERENT STAGES OF EMBRYONIC DEVELOPMENT

| AGE AT TREATMENT | NUMBER OF CULTURES | MINIATURE | MOSAIC | TOTAL | PER CENT MOSAICS |
|---|---|---|---|---|---|
| 0–24 hr. | 20 | 266 | 46 | 312 | 14.74 ± 1.35 |
| 3–4 days | 13 | 88 | 17 | 105 | 16.19 ± 2.43 |
| 6–7 days | 16 | 346 | 63 | 409 | 15.40 ± 1.20 |
| *Total treated* | *49* | *700* | *126* | *826* | *15.25 ± 0.84* |
| Control | 26 | 946 | 140 | 1086 | 12.89 ± 0.69 |

*Discussion.*—In spite of the fact that no significant increase in the rate of change in the treated material was observed the consistently higher rate in the three treated lots suggests a slight increase due to the treatment. If percentages for total treated and for control are taken as basis for calculation the treatment increased the rate of change about 1.83 per cent. The same treatment increases the general rate of mutations about one thousand per cent. Therefore the rate of increase in the unstable miniature-3 gene, if any, is so slight when compared with the rate of general increase that it could hardly be considered that the process responsible for the instability of the miniature-3 gene was affected by the treatment to the same extent as the processes responsible for changes in other genes.

Our position on this subject will be made clear after the following presentation of a part of our working hypothesis of the nature of the gene which is pertinent to this problem.

If a gene is visualized as a complex organic molecule composed of several molecular groups bound together (Demerec, 1933) an unstable gene may be pictured as such a molecule with one of the molecular groups chemically unstable. The change in that unstable molecular group takes place usually in one direction and the phenotypic result of that change is a reversion to wild-type. It is evident that a character determined by an unstable gene can be preserved through selection only.

The origin of an unstable gene may be visualized as a reversible chemical change in one of the molecular groups of the wild-type gene molecule and the origin of a stable gene may be pictured as a stable change in one of the molecular groups.

The whole complex of genes of an organism constitutes an interacting physiological system. When this system is balanced for a certain set of environmental conditions a wild-type phenotype is produced. Changes in a gene, which come about through changes in a molecular group of the gene molecule, may disturb the balance of the system to various degrees. The change in the phenotype is proportional to the degree of the disturbance caused in the balance of the system; thus if the disturbance is slight, the change in the gene may not register in the phenotype, and if the disturbance is great, the change in the gene may have a semilethal or a lethal effect. Between these two extremes many intermediate conditions are possible. Since all changes in a given locus usually affect the same characteristic, the degree of the effect would be proportional to the degree of the disturbance in the balance of the system. Thus a change in a gene molecule causing a slight disturbance in the system would produce a phenotype closer to wild-type than a change in a gene causing a greater disturbance in the system. Since the phenotype is determined by the degree of disturbance which the change in the gene molecule causes rather than by the nature of that change, chemically different changes in a gene may have identical phenotypic effects.

From this it would follow that from a wild-type a certain mutant may originate through the occurrence of one out of many possible changes in the gene molecule. On the other hand, in order that a reversion may occur it would be required for one certain chemical change to take place in a certain molecular group of the gene molecule. In case of stable genes this requirement may rarely or even may never be fulfilled and consequently reversions to wild-type, in case of stable genes, should be rare or absent. In case of unstable genes, however, the original chemical reaction which produced the change from the wild-type gene molecule to the mutant gene molecule is a reversible reaction. In that case, therefore, frequent reversions to the wild-type gene molecule are expected.

According to this view the action of x-rays on the mutation rate in general and on the reversion rate of unstable genes would be expected to

be different. The x-rays may well be expected to increase appreciably the rate of various changes in gene molecules, but they might not affect to an appreciable degree a rate of reaction which is already fairly high.

*Summary.*—The rate of reversions of the unstable miniature-3 gene has not been significantly increased by x-ray treatment of 600 to 1800 *r*-units.

Demerec, M., "Effect of Temperature on the Rate of Change of the Unstable Miniature-3 Gamma Gene of *Drosophila virilis*," these PROCEEDINGS, **18**, 430–434 (1932).

"What Is a Gene?" *Jour. Hered.*, **24**, 368–378 (1933).

"The Effect of X-Ray Dosage on Sterility and Number of Lethals in *Drosophila melanogaster*," these PROCEEDINGS, **19**, 1015–1020 (1933a).

---

## *EVIDENCE INDICATING THAT IN SCIARA THE SPERM REGULARLY TRANSMITS TWO SISTER SEX CHROMOSOMES*[1]

BY C. W. METZ

DEPARTMENT OF EMBRYOLOGY, CARNEGIE INSTITUTION OF WASHINGTON, AND DEPARTMENT OF ZOÖLOGY, JOHNS HOPKINS UNIVERSITY

Communicated December 15, 1933

One of the peculiar characteristics of chromosome behavior in *Sciara* is that exhibited by the "precocious" chromosome in the second spermatocyte division. As noted in earlier papers[2] this chromosome, instead of sending daughter halves to the two poles as the other chromosomes do, passes bodily to one pole in advance of the others. Consequently, one daughter cell (spermatid) receives both halves, or chromatids, of this chromosome, and the other daughter cell no representative of this kind. The latter cell is rudimentary, resembling a polar body, and subsequently degenerates. Thus all functional spermatids apparently receive the two chromatids in question. Owing to peculiarities of behavior in the primary spermatocyte (selective segregation and absence of synapsis)[3] it seems clear that the precocious chromosome represents a single parental chromosome of maternal origin and that in consequence the two chromatids under consideration are sisters, not homologues.

Before they move to the pole these two sister chromatids are presumably identical; but their peculiar subsequent behavior suggests, as noted previously,[4] that one may be significantly altered during the process of movement. This suggestion comes from the fact that, although still attached together as if at the equator, the two chromatids, even when close to the pole, possess spindle fibres extending toward opposite poles, and are obviously distorted because of the resultant opposing forces which tend to "pull" them apart. The constancy of this process throughout all the

species thus far examined[5] indicates that it must play a special rôle in reproduction here, and that its function may be to bring about a difference between the two chromatids so that when transmitted as separate chromosomes by the sperm they will no longer be identical.

It has been suspected for some time, as noted in an earlier paper,[4] that the "precocious" chromosome is a sex chromosome and that the above phenomena are concerned with sex determination. Evidence in support of this hypothesis has gradually accumulated until it now appears to be almost, if not entirely, conclusive. It is the purpose of the present paper to summarize this evidence and to note some of its implications in regard to problems of sex determination and chromosome behavior. The evidence shows that the two sister chromosomes derived from the "precocious" chromosome are transmitted by the sperms, and that an homologous chromosome is regularly transmitted by the eggs, giving three of this type in the fertilized egg. It also indicates strongly that one of the two sister chromosomes is functionally different from the other and that sex is finally determined during development in the embryo by a process of selective chromosome elimination which distinguishes between these two.

Since some of the most significant aspects of the present case have to do with changes in chromosome number, it is necessary at the outset to note certain features which bear on this topic but are not concerned with sex determination. As previously shown, the actual chromosome number varies considerably within the genus according to the species, to the stage of development and the sex of the individual, and to the nature of the tissue examined (references given below). There is, however, a basic uniformity in chromosome number when allowance is made for (1) alterations in number during development, and (2) the fact that two "kinds" of chromosomes are present. The latter include the "ordinary" chromosomes, with a basic number of eight (four pairs) in the mature germ-line and the female soma, and the peculiar, large "limited" chromosomes which are present only in the germ-line.[6] The "limited" chromosomes vary in number, when present, and in one species are not found at all[7] (at least in males, and hence almost certainly in females). Since they are evidently not involved in sex determination, as shown in the accounts referred to above, they may be ignored in the present summary. It is to be understood, therefore, that the chromosome numbers mentioned below include only the "ordinary" chromosomes. The actual total numbers may differ materially from these, except in the species having no "limited" chromosomes. The latter species has not yet been described taxonomically, hence it cannot be referred to here by name. Since much of the evidence in the present case has been derived from genetic studies, it should be observed that these involve primarily one species, *S. coprophila* Lint. This is also the species used most extensively for cytological observations.

The essential lines of evidence, aside from that already noted, may be summarized as follows:

1. The "precocious" chromosome is, of course, one of the "ordinary" chromosomes, not a "limited" chromosome. In each species thus far examined there is at least one other pair of ordinary chromosomes similar in size and shape to that to which the "precocious" belongs. For this reason the latter chromosome can only be identified directly at the second spermatocyte division, since in other mitoses it does not exhibit precocity in movement.

2. Maturation in the egg conforms to the usual type, with synapsis and random segregation of chromosomes, as shown by both cytological and genetic evidence.[7] The egg regularly transmits four "ordinary" chromosomes.

3. The sperm regularly transmits five ordinary chromosomes, two of which are the sister elements derived from the precocious chromosome. The latter two are homologous with one of the chromosomes in the egg. Together, therefore, these three form a set of three homologues in the fertilized egg. The other three chromosomes in the sperm are homologous, respectively, with the remaining three in the egg. Consequently the chromosome group of the fertilized egg includes three symmetrical pairs and a set of three, or nine in all, of the ordinary chromosomes. This group has been identified in the fertilized egg by DuBois,[8] whose observations I have carefully verified.

4. Later in development the cells of the germ-line in both sexes, and those of the soma in the female, possess only eight ordinary chromosomes, while the cells of the male soma possess only seven.[6] Fortunately, the paired association of the chromosomes here, as in other Diptera, makes it possible to distinguish the chromosome homologies accurately in the somatic cells. The female somatic group is composed of four symmetrical pairs and the male somatic group of three pairs and one unpaired chromosome. No set of three is present in either case.

5. The smaller number in these tissues results from a process of chromosome elimination during early development. The elimination from somatic nuclei in both sexes occurs during cleavage and has been carefully studied (DuBois[8] and unpublished evidence of the writer). The female soma loses one, and the male soma two entire chromosomes. In each case it is evident that the set of three is reduced by at least one member. Elimination in the germ-line apparently occurs later and has not yet been observed.

6. In interpreting the above evidence it is assumed that when a chromosome is eliminated it is different from the other chromosomes and is eliminated for that reason. The two chromosomes are eliminated from the male somatic nuclei at the same time (usually seventh or eighth cleavage),[8] and in at least one species they show distinctive characteristics even in the

prophase of the mitosis which leads to their elimination (unpublished evidence of the writer). From this it is concluded that the two have in common qualities not possessed by the other chromosomes. Such a conclusion, in turn, seems to require the assumption that they are homologous (a feature discussed more fully elsewhere). If this is true, one of them must be a derivative of the "precocious" chromosome and the other must be the other derivative of this chromosome or the homologue from the mother.

7. If it is granted that the two chromosomes under consideration are homologous with one another it must, obviously, be concluded that they are likewise homologous with the unpaired chromosome remaining in the soma (male). Since genetic evidence[9] seems to leave no doubt that the latter is the maternal $X$ chromosome, it follows that the two are also sex chromosomes and are paternal in origin, or in other words, that they are the two sister derivatives of the "precocious" chromosome.

8. In the case of the female soma, the paired association of the chromosomes, as well as their number, shows similarly that the chromosome eliminated during cleavage is either one of the two sister derivatives of the "precocious" chromosome or the maternal homologue of these. It must, therefore, on the present interpretation, be a sex chromosome, and since genetic evidence shows that the maternal $X$ chromosome is not eliminated, the eliminated one must be a derivative of the "precocious" chromosome.

This latter conclusion has special significance in that it requires the assumption that the two derivatives of the "precocious" chromosome are not identical at the time of elimination in the female. Otherwise there would be no distinction to account for the regular elimination of one and retention of the other. The mechanism by means of which such a distinction could be brought about has been noted above in the second paragraph.

9. In the case of the germ-line in the female, elimination involves superficially the same considerations as those applying to the female soma. Here again genetic evidence shows that the maternal $X$ chromosome is retained. On the present view, therefore, the eliminated one must be one of the two derivatives of the "precocious" chromosome. In this case, too, it must be assumed that the eliminated chromosome differs in some respect from its sister which is retained. In fact, the evidence all goes to show that the chromosome which is eliminated from the germ-line is identical with that eliminated from the soma.

These considerations raise the question as to which of the two derivatives of the "precocious" chromosome is eliminated from the male germline, and also as to what significance the whole mechanism has from the standpoint of sex determination. This question is considered more fully in another paper, but it may be observed here that in attempting to answer

it, and at the same time to harmonize the interpretation with the fact that in insects generally gonads as well as soma are self-differentiating, the following working hypothesis has been formulated:

It is postulated that at the second spermatocyte division the "precocious" chromosome, which appears always to be the maternal $X$, is so influenced as to alter one of its daughter halves by direct modification or by inactivation of part of its genic make-up. This results in a chromosome which no longer functions as an $X$, but behaves essentially like a $Y$ chromosome. In an egg (embryo) which is destined to produce a female this altered chromosome is eliminated from both soma and germ-line, with the result that the fly develops with the constitution $XX$ throughout, and is a female. In an embryo destined to produce a male both the altered and unaltered derivatives of the "precocious" chromosome are eliminated from the soma, which is thus $XO$ in constitution and develops the male characteristics. The germ-line of such an embryo, however, retains the altered instead of the unaltered, derivative of the "precocious" chromosome (together with the maternal $X$) and develops with the constitution $X"Y"$. This constitution is responsible for the production of testes.

The hypothesis may be modified in various ways. As it stands it assumes that soma and germ-line are both self-differentiating, as in other insects, and that the sex characteristics of each are determined by the sex chromosome constitution, after elimination has occurred. The initial processes of elimination are due to characteristics of the eggs at the beginning of development and these are in turn determined, in the material studied, by the sex chromosomes of the mothers—in which two types of $X$ chromosomes have been identified. This feature has been treated fully in earlier papers.[10] Whether elimination in the germ-line is independent of, or dependent on, that in the soma (which apparently precedes it) is as yet uncertain; but on the present interpretation the sex characteristics of soma and gonads are determined separately. If the former influence the latter it is only indirectly, through influencing chromosome elimination.

Whether or not the above interpretation is correct in detail the evidence indicates, as noted in earlier papers, that sex is finally determined in *Sciara* only after development of the embryo is under way, and that sex determination involves a series of processes, beginning, in some cases at least, before the egg is fertilized.

[1] This investigation has been aided by a grant from the NATIONAL RESEARCH COUNCIL, Committee for Research in Problems of Sex.

[2] Metz, Moses and Hoppe, *Zeits. ind. Abst-Vererb.*, **42**, 237 (1926); Metz, *Amer. Nat.*, **64**, 380 (1930).

[3] Metz, Moses and Hoppe, loc. cit.; Metz, these PROCEEDINGS, **12**, 690 (1926), and unpublished genetic evidence.

[4] Metz (1930), loc. cit.

[5] Ten or more species are included, some of which have not been identified taxonomically. They include representatives of both sections of the genus.

[6] Metz, *Biol. Zentralb.*, **51**, 121 (1931); Metz and Schmuck, these PROCEEDINGS, **17**, 272 (1931).

[7] Metz, *Amer. Nat.*, **68**, 487 (1929).

[8] DuBois, these PROCEEDINGS, **18**, 352 (1932); *Jour. Morph.*, **54**, 161 (1932); *Zeits. f. Zelf. u. mik. Anat.*, **19**, 395 (1933).

[9] Metz and Schmuck, *Genetics*, **16**, 225 (1931).

[10] Metz and Schmuck, these PROCEEDINGS, **15**, 863 (1929); *Ibid.*, **15**, 867 (1929); Metz, *Quart. Rev. Biol.*, **6**, 306 (1931).

---

## *ON THE OCCURRENCE OF TRANSLOCATIONS AND AUTOSOMAL NON-DISJUNCTION IN DROSOPHILA MELANOGASTER*

BY CURT STERN[1]

DEPARTMENT OF ZOÖLOGY, UNIVERSITY OF ROCHESTER, ROCHESTER, N. Y.

Communicated November 28, 1933

Recent findings on different plants indicate that the presence of extra chromosomes in an organism facilitates the occurrence of exchanges of non-homologous chromosomal parts. It seemed desirable to test if such exchanges occur also in heteroploid *Drosophila melanogaster* both from the point of the validity of the generalization of the findings as well as for the possibility of finding a method to obtain non-homologous exchanges in non-radiated flies. Triploid Drosophilae were chosen for this investigation. Translocations within one chromosomal pair as well as between non-homologous chromosomes were to be expected if the findings on plants would hold true for *Drosophila*. However, as a first part of the investigation, only the possibility of translocations between non-homologous chromosomes was tested.

The outcome was fully negative: None of the more than 900 haploid gametes coming from triploid females contained a translocation between the first, second or third chromosomes (the fourth chromosome was—on account of its comparatively small size—expected to be of less importance than the others and its fate therefore not followed). The result can be explained as due to the fact that the frequency of translocations between non-homologous chromosomes in triploid Drosophilae is very low. If one considers the fact, however, that the production of a haploid egg from triploid females, which contains the two non-homologous chromosomes that underwent a possible interchange, is only one of different possibilities during meiosis, the number of 900 gametes tested might still have been too low. Certainly the method used has proved to be of no value for obtaining translocations.

A brief record of the data, which constitute not only the basis for the

above considerations, but are also the most extensive on the possibility of translocations of the kind in question in non-radiated diploid males, follows below. Furthermore, some new facts about chromosome distribution during reduction in special types of males and in triploid females came to light, showing the occurrence of autosomal non-disjunction.

Triploid females were mated to males carrying in different chromosomes the dominant characters Curly (wing; Cy, II), Plum (eye; Pm, II), Stubble (bristles; Sb; III) and Hairless (H; III):

$$3N\,♀ \times \frac{Cy}{Pm}\,\frac{Sb\ C}{H}\,♂.$$

Both Cy and Pm are associated with inversions which suppress practically all crossing-over between these two genes. Sb was also associated with a crossover suppressor (C), so that no crossing-over occurs between Sb and H. Some of the triploid females contained different recessive genes in the X-chromosome. Possibly some also contained H in single dose, as appeared later.

TABLE 1

$$3N\,♀ \times \frac{Cy}{Pm}\,\frac{Sb\ C}{H}\,♂$$

$$\frac{+^{Cy}}{+^{Pm}}\,\frac{+^{Sb}}{+^{H}}\,♀♀ \times F_1♂$$

| Types of $F_1$—♂♂ | Cy Sb | Cy H | Pm Sb | Pm H | Total |
|---|---|---|---|---|---|
| Number of each type tested (fertile) | 229 | 219 | 228 | 227 | 903 |

The male offspring of this cross consisted of four different regular classes in about equal numbers in regard to the four dominant autosomal genes: $\frac{Cy}{+}\,\frac{Sb}{+}$, $\frac{Cy}{+}\,\frac{H}{+}$, $\frac{Pm}{+}\,\frac{Sb}{+}$ and $\frac{Pm}{+}\,\frac{H}{+}$. A total of 1172 such males were mated individually to females carrying the wild type allels of these dominant characters. 269 cultures were sterile (in large proportion due to unfavorable conditions in one batch), leaving 903 fertile matings. Had the P-males of these matings, being the $F_1$ of the triploid cross, contained any translocation between two or three of the chromosomes I–III, then an apparent linkage between the characters carried by the chromosomes involved in the translocations (sex–Cy or Pm–Sb or H), would have appeared. The findings that free recombinations between all characters occurred in all 903 cultures is evidence that no translocation occurred either in the gametes from the 3N-female or in those of their diploid mates (table 1).

Besides the four regular classes eight other types of males appeared. *They contained either none or both of the two dominant genes carried in the two chromosomes II or in the two chromosomes III of their father* (table 2).

TABLE 2

| Non-disjunction of II | | | | Non-disjunction of III | | | |
|---|---|---|---|---|---|---|---|
| Both II from ♀ No II from ♂ | | Both II from ♂ No II from ♀ | | Both III from ♀ No III from ♂ | | Both III from ♂ No III from ♀ | |
| + Sb / + + | + H / + + | Pm Sb / Cy + | Pm H / Cy + | Pm + / + + | Cy + / + + | Pm Sb / + H | Cy Sb / + + |
| 5 | 1 | 6 | 1 | 1 | 1 | 4 | 1 |
| 6 | | 7 | | 2 | | 5 | |

A test, carried out in each case, showed that the *two* mutant genes of one original pair were carried in different chromosomes in these exceptional males and were not, as a result of crossing-over in an earlier generation, carried together in one chromosome. This test was not possible in case *none* of the genes was present in the $F_1$-males, but the possibility of not otherwise detected crossing-over was very low. The fact that the *P*-males transmitted to these sons none or both of either their chromosomes II or III is best explained by the assumption of non-disjunction of these autosomes in the male. This non-disjunction must be comparatively frequent, considering, for example, the finding of 13 such cases for chromosome II among 903 gametes. This is especially clear if we consider that the appearance of the exceptional males is only possible when such non-disjunctional spermatozoa meet the right type of egg to give a viable male combination. For the spermatozoa which carry no chromosome II or III the chances of finding a favorable egg are good as a triploid female produces normally many eggs with two chromosomes II or III. For the spermatozoa with two chromosomes II or III, however, only eggs with no chromosome II or III will be fitted. Such eggs were not known to occur, but the fact that 7 exceptional $F_1$-males carrying both Cy and Pm in separate chromosomes II were found seems to indicate their occurrence. It may be added that three of these males (in the $\frac{Cy}{Pm}\frac{Sb}{+}$ class) carried also Bar (eye; B, I), which came from the 3N mother thus excluding any possibility of contamination. For the classes carrying both Sb and H the evidence is not quite as clear, as there was the possibility of H being present in some of the 3N-females, thus accounting for this class without the assumption of non-disjunction. The proportion in which the eight exceptional classes appeared are striking in different respects but should be tested by experiments especially designed for it before a discussion can be of value. Such experiments which will throw more light on the

new findings of non-disjunction in autosomes of triploid females and of males carrying inversions in their autosomes seem to be desirable. It should be noted that cases of autosomal non-disjunction in males which carry inversions have already been found by *Dobzhansky*[2] (1933) and by *J. Schultz* (unpublished). *Dobzhansky* and *Darlington* have observed cases of non-disjunction of autosomes cytologically in *Drosophila pseudo-obscura* (unpublished).

*Summary.*—1. No translocations between chromosomes I, II and III of *Drosophila melanogaster* were found in more than 900 eggs from triploid females and the equal number of spermatozoa from diploid males.

2. Non-disjunction of the autosomes in males, resulting in spermatozoa with two or none of a special pair of autosomes was frequent.

3. Triploid females frequently produced eggs which did not receive any of the chromosomes II or III.

For laboratory facilities and for help and advice I am thankful to Drs. T. H. Morgan, C. W. Beadle, Th. Dobzhansky, Barbara McClintock and J. Schultz.

[1] Work done—as Fellow of the Rockefeller Foundation—at the Biological Laboratories of the California Institute of Technology, Pasadena, Cal.

[2] Dobzhansky, Th., *Zeitschr. ind. Abstgsl.*, **64**, 269–309 (1933).

---

## *UNISEXUAL MAIZE PLANTS AND THEIR RELATION TO DIOECISM IN OTHER ORGANISMS*

BY DONALD F. JONES

CONNECTICUT AGRICULTURAL EXPERIMENT STATION, NEW HAVEN, CONNECTICUT

Communicated November 29, 1933

Dioecious plants occur naturally in at least 61 families scattered among some 48 orders of angiosperms in both monocotyledons and dicotyledons. With these unisexual plants there are often monoecious and hermaphroditic individuals in the same species. From the widespread distribution in the plant kingdom, it seems probable that various systems of partial and complete sex-separation have been developed. In 7 dioecious species of bryophytes and 51 dioecious species of angiosperms there are differences in the size, shape or number of chromosomes associated with the two sexes. In all but a few cases the female gametes are alike. In 33 dioecious species no allosomes have been found. Many hypothetical factors, both germinal and environmental have been postulated to account for the appearance of different sexes in the offspring from the same parents.

The occurrence of dioecious forms in a species that heretofore has not

shown them and their continued propagation through four generations gives an actual illustration of the working of specific germinal factors resulting in sex-separation.

Dioecious maize has been developed by combining two recessive genes, tassel-seed-2 and silkless, located on different chromosomes, now numbered I and II. Plants with these two recessive genes in the homozygous condition have the staminate flowers in the terminal inflorescences replaced with carpellate flowers. Functionally these plants are females. The male plants have the same factorial composition except that the tassel seed gene is heterozygous. The terminal flowers are staminate as in normal maize. The lateral inflorescences are entirely sterile. In this way the seed parents produce only seed and the pollen parents only pollen.

Fourteen dioecious families in all have been grown. Six of these have been continued through the fourth generation. In a total of 967 plants in all families, 390 are males and 515 females, with 61 normals. Twenty-one progenies produced only male and female plants; of these, 170 are males and 210 females, a ratio of 1 to 1.24.

Evidence from several sources proves that the female plants are homozygous, $ts_2\ ts_2\ sk\ sk$, in composition, and that the male plants are heterozygous, $Ts_2\ ts_2\ sk\ sk$. The male plants are stable in their sex expression in all environments in which they have been grown. Abortive pistils have been noted on a few plants but no seeds have been produced. In the female, the terminal inflorescences are highly variable ranging from all carpellate flowers with no stamens to a large number of flowers with functioning stamens and only a few seeds. This variation is partly developmental and partly genetic. The latter is indicated by the fact that persistent differences in this respect are seen in the several families all grown under the same conditions.

The tassel-seed-2 gene not only suppresses the staminate flowers in the terminal inflorescences but also forces normally aborted ovules into full growth. This results in the crowded and irregular kernel arrangement in the lateral inflorescences that are characteristic of tassel seed plants. For these reasons tassel-seed-2 is considered to be a female-promoting gene. Other germinal factors of this general type are known in maize.

The silkless factor suppressing the carpellate flowers in all parts of the plant is considered as a male-promoting gene. Other genes having a similar action have been located in several different chromosomes of maize. When homozygous tassel-seed-2 and silkless are together in the same individual the silkless gene has no apparent effect.

This dioecious maize has originated from normal plants showing no irregularities in chromosome number or structure. The enforced heterozygosity in one sex may permit a viable deficiency to persist in one member of the heterozygous pair if it should occur. This would result in chromosome

inequality of the two sexes. Inversions in this chromosome pair would reduce or suppress crossing-over in one sex. An accumulation of recessive factors in the protected chromosome, interacting with complementary factors in other chromosomes, could bring about an intensification of the two sex types together with secondary sexual characters. At least the interaction of two specific loci holds the balance of control over the other germinal factors and over the environmental influences and supplies a working mechanism whereby separate-sexed individuals are produced and maintained from generation to generation.

The evidence supporting these brief statements will be published in *Genetics*.

---

## *CONCERNING COMPACT CONTINUA WHICH CONTAIN NO CONTINUUM THAT SEPARATES THE PLANE*

BY R. L. MOORE

DEPARTMENT OF PURE MATHEMATICS, UNIVERSITY OF TEXAS

Communicated November 17, 1933

It has been shown, by Mazurkiewicz,[1] that every arc which lies in a compact acyclic continuous curve $M$ is a subset of an arc which lies in $M$ but which is itself not a proper subset of any other such arc. In the present paper it will be shown that if, in a plane $\alpha$, $S$ is a compact continuum containing no continuum that separates $\alpha$ and $K$ is a subcontinuum of $S$ which is irreducible between some two points, then $K$ is a subset of some subcontinuum of $S$ which is irreducible between some two points but which is itself not a proper subset of any other such subcontinuum of $S$. With the aid of this result it will be proved that if, in the plane $\alpha$, $M$ is a non-degenerate, compact and atriodic continuum which does not separate $\alpha$, then $M$ is irreducible between some two points.

Let $S$ denote a definite compact continuum lying in the plane $\alpha$ and containing no continuum that separates $\alpha$. If $X$ and $Y$ are any two points of $S$ the notation $XY$ will be employed to designate the[2] irreducible subcontinuum of $S$ from $X$ to $Y$.

THEOREM 1. *If, in $S$, the indecomposable continuum $M$ contains the point $B$ of the irreducible continuum $AB$ and $M$ is irreducible from $B$ to some point $O$ not lying on $AB$, then the continuum $M + AB$ is irreducible from $A$ to $O$.*

*Proof.*—Since no subcontinuum of $S$ separates the plane $\alpha$, $M \cdot (AB)$, the common part of $M$ and $AB$, is[2] a continuum, and so is $M \cdot (AO)$. But $AO$ is necessarily a subset of $M + AB$ and therefore it contains a point of

$M\cdot(AB)$. It follows that $M\cdot(AB) + M\cdot(AO)$ is a subcontinuum of $M$. But it contains both $B$ and $O$. Therefore $M\cdot(AB) + M\cdot(AO) = M$. But $M$ is indecomposable and $M\cdot(AB)$ is a proper subcontinuum of $M$. Hence $M\cdot(AO)$ is identical with $M$. Hence $B$, as well as $A$, belongs to $AO$ and therefore $AB$, as well as $M$, is a subset of $AO$ and consequently $M + AB$ is identical with the irreducible continuum $AO$.

THEOREM 2. *If, in S, AO is identical with AY and O does not belong to XA and XA + AY = XY, then XY = XO.*

*Proof.*—Case 1. Suppose $OY$ does not contain $A$. The continuum $XO + OY$ is identical with $XY$ and therefore contains $A$, and therefore $XO$ contains $A$, and $XA + AO$ is identical with $XO$ and thus $XY$ is irreducible from $X$ to $O$.

Case 2. Suppose $OY$ does contain $A$. If $Y = O$, $XY = XO$. Suppose $Y$ is distinct from $O$. Then $AO$ is irreducible between each two of the three distinct points $A$, $O$ and $Y$ and therefore, by a theorem due to Janiszewski and Kuratowski,[3] it is an indecomposable[4] continuum. Hence, by Theorem 1, $XY$ is irreducible from $X$ to $O$.

THEOREM 3. *If AB is an irreducible subcontinuum of S between the two distinct points A and B, there exists a point O such that AO contains AB but such that AO is not a proper subset of any irreducible subcontinuum of S from A to another point.*

*Proof.*—Let $G$ denote the set of all subcontinua $x$ of $S$ such that $x$ contains $AB$ and is an irreducible continuum from $A$ to some other point of $S$. Suppose that every continuum of the set $G$ is a proper subset of some other continuum of $G$. There exists a well-ordered sequence $\alpha$ whose terms are the elements of $G$. There exists a well-ordered subsequence $\beta$ of $\alpha$ such that (1) the first term of $\beta$ is the first term of $\alpha$, (2) if $\gamma$ is a subsequence of $\beta$ and there exists a continuum belonging to $\alpha$ and containing all the continua of the sequence $\gamma$ then the first such continuum in $\alpha$ is the first term of $\beta$ that, in $\beta$, follows all the terms of $\gamma$, (3) if $\gamma$ is a subsequence of $\beta$ such that no continuum of the sequence $\alpha$ contains every continuum that belongs to $\gamma$ then there is no term of $\beta$ following, in the sequence $\beta$, all the terms of $\gamma$. By a theorem[5] of Sierpinski's, the sequence $\beta$ is countable. It follows that there exists a subsequence $t_1, t_2, t_3, \ldots$ of $\beta$ such that (1) for each $n$, the continuum $t_n$ is a proper subset of $t_{n+1}$, (2) if $x$ is any element of $\beta$ then there exists a natural number $n$ such that $t_n$ follows $x$ in $\beta$. Let $Q$ denote the sequence $t_1, t_2, t_3, \ldots$. No continuum of the set $G$ contains every continuum of the sequence $Q$. For each $n$, $t_n$ is an irreducible continuum from $A$ to some point $A_n$, and $A_{n+1}$ belongs to $S - t_n$.

Case 1. Suppose there exist a number $m$ and an infinite ascending sequence of numbers $n_1, n_2, n_3, \ldots$ such that, for each $i$, $A_m$ belongs to $A_{n_i} A_{n_{i+2}}$. In this case, for each $i$, let $B_i$ denote $A_{n_{i+2}}$, let $H$ denote

the point set $B_1B_2 + B_2B_3 + B_3B_4 + \ldots$ and let $\overline{H}$ denote the continuum consisting of $H$ together with all of its limit points. For each $n$, $AB_{n+1} = AA_m + A_mB_{n+1}$; and $AA_m$ is a subset of $AA_{n_2}$ and therefore $AA_m$ cannot contain $B_n$. Hence $A_mB_{n+1}$ contains $B_n$. But $B_{n+1}B_{n+2}$ contains $A_m$ and $B_{n+1}$. Hence $B_{n+1}B_{n+2}$ contains $A_mB_{n+1}$ and therefore $B_n$. Hence it contains $B_nB_{n+1}$. If $i$ and $j$ are both less than $n$ then $B_iB_n = B_jB_n$.

Suppose now that $L$ is any proper subcontinuum of $\overline{H}$. If $L$ contains no point of $H$ it is clearly a continuum of condensation of $\overline{H}$. Suppose $L$ contains a point of $H$. Since $L\cdot H = L\cdot(B_1B_2) + L\cdot(B_1B_3) + L\cdot(B_1B_4) + \ldots$, there exists a number $m$ such that $L\cdot(B_1B_m)$ is a non-vacuous proper subset of $B_1B_m$. Suppose $n > m$. Let $T$ denote the set of all points $X$ such that $B_nB_{n+1}$ is irreducible from $B_{n+1}$ to $X$. Suppose first that the continuum $B_1B_n$ is not indecomposable. Then $T$ contains $B_1B_n$. Hence the continuum $L\cdot(B_nB_{n+1})$ does not contain all of $T$. But $T$ is a subset of every subcontinuum of $B_nB_{n+1}$ that contains both a point of $T$ and a point not belonging to $T$. Therefore $L$ contains no point of $B_nB_{n+1} - T$. But[6] every point of $T$ is a limit point of $B_nB_{n+1} - T$. Hence every point of $L\cdot(B_1B_n)$ is a limit point of $\overline{H} - L$. Suppose, on the other hand, that $B_1B_n$ is indecomposable. Then every point of the continuum $L\cdot(B_1B_n)$ is a limit point of $B_1B_n - L\cdot(B_1B_n)$. Thus, in either case, if $n > m$ every point of $L\cdot(B_1B_n)$ is a limit point of $\overline{H} - L$. Therefore every point of $L\cdot H$ is a limit point of $\overline{H} - L$. Suppose now $P$ is a point of $L\cdot\overline{H} - L\cdot H$. The point $P$ is a limit point of $H$ and therefore either of $H - L\cdot H$ or of $L\cdot H$. In the latter case, since every point of $L\cdot H$ is a limit point of $\overline{H} - L$, $P$ is a limit point of $\overline{H} - L$. It follows that $L$ is a continuum of condensation of $\overline{H}$. Therefore by a theorem due to Mazurkiewicz[1] $\overline{H}$ is an indecomposable continuum.

Let $N$ denote the continuum $AB_1 + \overline{H}$. There exists a point $O$ such that $\overline{H} = B_1O$. Since $\overline{H}$ is not a subset of $AB_1$, $O$ does not belong to $AB_1$. Hence, by Theorem 1, $N$ is irreducible from $A$ to $O$.

Case 2. Suppose there exists no number $m$ such that, for some ascending infinite sequence of numbers $n_1, n_2, n_3, \ldots$, the point $A_m$ belongs to $A_{n_i}A_{n_i+1}$ for every $i$. Then there exists an infinite sequence of numbers $m_1, m_2, m_3, \ldots$ such that if $j > i + 1$ then $A_{m_i}$ does not belong to $A_{m_i+1}\,A_{m_j}$. For each $n$, let $B_n$ denote $A_{m_n}$ and let $Q_n$ denote $B_nB_{n+1} + B_{n+1}B_{n+2} + B_{n+2}B_{n+3} + \ldots$. This point set is identical with $B_nB_{n+1} + B_nB_{n+2} + B_nB_{n+3} + \ldots$. If, for some number $m$ and some number $j$ greater than $m + 2$, $AB_m$ and $B_{m+2}\,B_j$ had a point in common, then $AB_m + B_{m+2}\,B_j$ would be a subcontinuum of $S$ containing $A$ and $B_{m+2}$ and therefore $AB_{m+2}$ and consequently the point $B_{m+1}$ which is impossible since $B_{m+1}$ belongs neither to $AB_m$ nor to $B_{m+2}B_j$. Hence $AB_n$ and $Q_{n+2}$ have no point in common.

Suppose that there exists a number $k$ such that $B_k$ is a limit point of the point set $Q_{k+1}$. For each $i$, the continuum $\bar{Q}_{k+i}$ contains $B_k$ and $B_{k+i}$ and therefore $B_kB_{k+i}$. Hence, for each $i$, $\bar{Q}_{k+i} = \bar{Q}_k$. Suppose now that $L$ is a proper subcontinuum of $\bar{Q}_k$. If $L$ contains the points $B_i$ and $B_j$ it contains $B_iB_j$. Hence if there exists an infinite ascending sequence of numbers $n_1, n_2, n_3, \ldots$, all greater than $k$, such that the points $B_{n_1}, B_{n_2}, B_{n_3}, \ldots$ all belong to $L$, then $L$ contains $Q_{n1}$ and therefore $\bar{Q}_{n1}$ which is identical with $\bar{Q}_k$. This involves a contradiction. Hence there exists a number $m$ such that $L \cdot Q_k$, if it exists, is a subset of $B_kB_{k+m}$. The point sets $B_kB_{k+m}$ and $Q_{k+m+2}$ are mutually exclusive. Hence $Q_{k+m+2}$ is a subset of $\bar{Q}_k - L$. But $\bar{Q}_{k+m+2} = \bar{Q}_k$. Therefore $L$ is a continuum of condensation of $\bar{Q}_k$. Hence $\bar{Q}_k$ is[1] an indecomposable continuum. The continuum $\bar{Q}_k + AB_k$ is irreducible from $A$ to some point $O$ of $\bar{Q}_k - Q_k$.

Suppose there exists no number $k$ such that $B_k$ is a limit point of $Q_{k+1}$. Let $W$ denote the point set $AB_1 + B_1B_2 + B_2B_3 + \ldots$, let $N$ denote $\overline{W}$ and let $O$ denote a limit point of the point set $B_1 + B_2 + B_3 + \ldots$. Suppose that $O$ belongs to $AB_n$ for some $n$. Then $AB_n + \bar{Q}_{n+2}$ is a continuum containing $A$ and $B_{n+2}$, and therefore containing the irreducible continuum $AB_{n+2}$ which contains $B_{n+1}$. But $B_{n+1}$ belongs neither to $AB_n$ nor to $Q_{n+2}$. Therefore $B_{n+1}$ is a limit point of $Q_{n+2}$. This involves a contradiction. Hence $O$ does not belong to $W$. The continuum $AO$ is a subset of $N$ and, for each $n$, $AO = (AO) \cdot (AB_{n+1}) + (AO) \cdot (B_{n+1}\, B_{n+3}) + (AO) \cdot \bar{Q}_{n+3}$. But $AB_{n+1}$ and $\bar{Q}_{n+3}$ are mutually exclusive closed point sets. Hence, since $AO$ is a continuum $(AO) \cdot (B_{n+1}\, B_{n+3})$ is non-vacuous. It follows that, for each $n$, $AO$ contains $Bn$ and therefore $AB_n$. Hence it contains $W$. It follows that $AO$ is identical with $N$.

Both in Case 1 and in Case 2, $AO$ is an irreducible continuum from $A$ to $O$ lying in $S$ and containing every continuum of the sequence $t_1, t_2, t_3, \ldots$. Thus the supposition that Theorem 3 is false leads to a contradiction.

THEOREM 4. *Under the hypothesis of Theorem* 3, *if AO satisfies the conditions of the conclusion of that theorem, then if XY is an irreducible subcontinuum of S containing AO, XY is irreducible either from O to X or from O to Y.*

*Proof.*—Suppose, on the contrary, that neither $XO$ nor $YO$ is identical with $XY$. Since $AX + AY = XY$, $O$ belongs either to $AX$ or to $AY$. Suppose it belongs to $AY$. Then $AY$ is identical with $AO$. Hence $AX + AO = XY$.

Case 1. Suppose $O$ belongs also to $AX$. Then $AX$ is identical with $AO$ and consequently with $AY$. Therefore $XY = AX + AX = AX$. Hence,[3] in this case, $XY$ is an indecomposable continuum. Since $XY$ is irreducible from $X$ to $Y$, $X$ and $Y$ do not both belong to the composant of $XY$ that contains $O$. Hence $XY$ is identical either with $OX$ or with $OY$.

Case 2. Suppose $O$ does not belong to $AX$. Then, by Theorem 2, $XY$ is irreducible from $X$ to $O$.

An analogous argument applies in case $O$ belongs to $AX$.

THEOREM 5. *If $AB$ is an irreducible subcontinuum of $S$ between the two points $A$ and $B$ then there exists an irreducible subcontinuum of $S$ which contains $AB$ but which is not a subset of any other irreducible subcontinuum of $S$.*

*Proof.*—By Theorem 3, there exists a point $O$ of $S$ such that $AO$ contains $AB$ but is not a proper subset of $AX$ for any point $X$ of $S$. By the same theorem there exists a point $C$ of $S$ such that $OC$ contains $AO$ but is not a proper subset of $OZ$ for any point $Z$ of $S$. Suppose now that there exist two points $X$ and $Y$ belonging to $S$ and such that $OC$ is a proper subset of $XY$. Then, by Theorem 4, since $AO$ is a proper subset of $XY$, $XY$ is identical either with $OX$ or with $OY$. But this involves a contradiction. The truth of Theorem 5 is therefore established.

THEOREM 6. *In order that $S$ should be triodic[7] it is necessary and sufficient that there should exist three points of $S$ such that no one of them separates the remaining two from each other in $S$ in the weak sense.*

THEOREM 7. *If $S$ is non-degenerate and atriodic then it is irreducible between some of its points.*

*Proof.*—By hypothesis, $S$ contains two distinct points $A$ and $B$. There exists an irreducible continuum $AB$. By Theorem 5, there exist two points $O$ and $C$ belonging to $S$ and such that $OC$ contains $AB$ but is not itself contained in any other irreducible subcontinuum of $S$. Suppose now that there exists a point $X$ of $S$ not lying on $OC$. Then $OC$ does not contain $X$, $OX$ does not contain $C$, and $CX$ does not contain $O$. But this is contrary to Theorem 6.

[1] Mazurkiewicz, S., *Fund. Math.*, **2**, 119–130 (1921).

[2] Janiszewski has shown that if $X$ and $Y$ are two points of a compact continuum $M$ in any one of a certain class of spaces which includes all metric spaces then there is at least one irreducible subcontinuum of $M$ from $X$ to $Y$. Cf. Janiszewski, S., *J. éc. polytech.*, 2° Série, 79–170 (1912). Since no subcontinuum of $S$ separates the plane $\alpha$, therefore, by another theorem of Janiszewski's, the common part of every two intersecting subcontinua of $S$ is itself a continuum; see his paper "Sur les coupures du plan faites par des continus," *Prace mat.-fizycne*, **26**, 11–63 (1913). It follows that if $X$ and $Y$ are two points of $S$ there is *only one* irreducible subcontinuum of $S$ from $X$ to $Y$. Cf. Moore, R. L., these PROCEEDINGS, **12**, 361 (1926).

[3] S. Janiszewski and C. Kuratowski, *Fund. Math.*, **1**, 212 (1920).

[4] A continuum is said to be indecomposable if it is not the sum of two continua neither of which is identical with it. For a proof that such continua exist, cf. Brouwer, L. E. J., *Math. Ann.*, **68** (1910). The term indecomposable was introduced by Mazurkiewicz, *Fund. Math.*, **1**, 35–39 (1920).

[5] Sierpinski, W., *Fund. Math.*, **2**, 179–188 (1921).

[6] Kuratowski, K., *Ibid.*, **10**, 236 (1927).

[7] Moore, R. L., *Ibid.*, **13**, 262 (1929).

# DOES INSTABILITY IMPLY TRANSITIVITY?

BY MARSTON MORSE

DEPARTMENT OF MATHEMATICS, HARVARD UNIVERSITY

Communicated December 1, 1933

Much has been written of late concerning the hypotheses of metric transitivity. One can refer for example to the historical summary[1] by Birkhoff and Koopman. Here references are given to the contributions of various writers including the recent important papers by Birkhoff, Hopf, Koopman, v. Neumann and P. Smith. However, it has not yet been proved that there is a distinction in the analytic case between metric and regional transitivity. Outside of very simple examples of integrable motion or motion on a surface of negative curvature,[4] the work prior to the present paper has not shown that metric or regional transitivity exists in general analytic cases. The present paper partially fills this gap with a general theorem on the existence of regional transitivity.

A dynamical system is said to be regionally transitive if there exists at least one motion $g$ whose closure in phase space is identical with the phase space. In such a case we term $g$ a *completely transitive motion*. This paper is concerned with geodesic motion on a generalized surface $R^*$ of genus $p > 1$. We shall introduce a conception of *uniform instability* of the geodesics on $R^*$, and show that uniform instability of the geodesics on $R^*$ implies regional transitivity. Surfaces on which the geodesics are uniformly unstable include all surfaces of negative curvature, as well as surfaces which contain large areas of positive curvature. Professor Birkhoff has made the very interesting conjecture that regional transitivity is implied by the condition that all closed motions be unstable and of minimum type. We are unable to confirm this opinion. Such evidence as we possess seems to be slightly against the conjecture. The present result of the author is the first in which a hypothesis on the separate geodesics leads to a general conclusion as to transitivity.

We shall say that a motion $g$ is transitive relative to a class $\Omega$ of motions in phase space if the closure of $g$ includes $\Omega$ as a subset. As an important example we consider a fundamental class $A$ of minimizing geodesics on $R^*$ and show that there always exist geodesics transitive to the motions defined by $A$. The class $A$ is so chosen that it contains all motions on $R^*$ when the system is uniformly unstable. We also state a general theorem on the deferring of transitivity in a regionally transitive system.

We suppose that we have a Fucksian group $G$ of non-singular fractional linear transformations of the plane $z = u + iv$ carrying the interior $S$ of the unit circle into $S$. We suppose all members of the group are of hyperbolic type with fixed points on the unit circle, and that the group has a

finite number of generators. We regard $S$ as a hyperbolic plane of non-euclidean geometry with the circles orthogonal to the unit circle representing the $H$-straight lines ($H$ is written for hyperbolic). The $H$-length may be given by the integral

$$\int \frac{d\sigma}{1-u^2-v^2} \tag{1}$$

where $d\sigma$ is the euclidean differential of arc length. We assume that there exists a fundamental domain for the group $G$ in the form of a region $S_0$ bounded by a finite set of $H$-straight lines and that the closure of $S_0$ has no points on the unit circle. Certain of the boundary points of $S_0$ belong to $S_0$. We shall not give the details here.

Instead of assigning $S$ the $H$-metric (1) we can also assign $S$ an $R$-metric ($R$ is written for Riemannian) given by a positive definite form

$$ds^2 = E(u,v)du^2 + 2F(u,v)dudv + G(u,v)dv^2 \tag{2}$$

with coefficients of class $C^3$. We suppose the form (2) is invariant under $G$ and defines a Riemannian manifold $R$. We can, if we choose, regard congruent points on the boundary of $S_0$ as identical. So regarded, $S_0$ taken with the metric (2), will define a generalized closed surface $R^*$. The genus of $R^*$ will be an integer $p > 1$.

To any two points on $S$ we assign an $R$-distance equal to the minimum length of $R$-geodesics joining the two points. A point $P$ on $S$ with a direction $\theta$ at $P$, measured in the usual way in the $(u, v)$ plane, will define an *element* $E$ on $S$. If $E'$ and $E''$ are any two elements on $S$ with directions represented by $\theta'$ and $\theta''$, the $R$-angle between $E'$ and $E''$ will be defined as the minimum of

$$|\theta' - \theta'' + 2n\pi|$$

for all integers $n$, positive, negative or zero. Two elements $E'$ and $E''$ on $S$ will be said to have an $R$-distance which is the sum of the $R$-distance between their initial points and their $R$-angle. The elements congruent to $E'$ and $E''$ on $S_0$ will be said to have an $R^*$-distance which is the greatest lower bound of the $R$-distance between all pairs of elements which are respectively congruent to $E'$ and $E''$.

The present paper is based on an earlier paper by the writer.[2] We take over the terminology of the earlier paper. We showed in this paper that there always exist two "boundary" geodesics on $R$ of class $A$ of the type of any given $H$-straight line.

*By the type hypothesis on R we here mean the hypothesis that any two boundary geodesics of the same type are identical.*

In the earlier paper we showed that the boundary geodesics were limit geodesics of the set of periodic boundary geodesics. With the aid of this fact we can readily establish the following theorem.

THEOREM 1. *If the type hypothesis holds the geodesics on $R^*$ which are completely transitive have the power of the continuum.*

In case $R^*$ is a surface of constant negative curvature Hedlund[3] has shown that the above geodesics define motions in phase space (the space of their elements) which have the measure of the space. No such theorem has been proved for the present general case. Theorem 1 was announced by the author in February, 1932, at the Harvard Colloquium.

Let $M$ denote the phase space corresponding to geodesic motion on $R^*$, that is, the ensemble of all elements tangent to geodesics on $R^*$ with the distance between elements in phase space defined as previously. The proof of Theorem 1 leads to Theorem 2.

THEOREM 2. *If the type hypothesis holds and $h$ is any arc in the space $M$ such that the motions on $M$ which intersect $h$ define more than one geodesic on $R^*$, then the completely transitive motions which intersect $h$ have the power of the continuum.*

*Uniform Instability.*—The two preceding theorems show the importance of the type hypothesis. We shall give a sufficient condition for this hypothesis to hold.

To that end let $g$ be any geodesic belonging to $R$. We refer $R$ to normal geodesic coördinates $(x, y)$ such that neighboring $g$

$$ds^2 = C^2(xy)dx^2 + dy^2 \tag{3}$$

where $y = 0$ along $g$ and $C(x, 0) = 1$. The parametric curves, $x$ constant, are geodesics normal to $g$, and $y$ is the arc length along these geodesics measured from $g$. The equation of normal variation from $g$ takes the form

$$\frac{d^2w}{dx^2} + K(x)w = 0 \tag{4}$$

where $K(x)$ is the curvature of $R$ at the point $x$ on $g$. We say that the differential equation (4) and its solutions are *based* on $g$. Let $w(x)$ be any solution of (4) such that

$$w^2(o) + w'^2(o) = 1. \tag{5}$$

We shall say that the geodesic system on $R^*$ is *uniformly unstable* if the following two conditions are satisfied by every geodesic $g$ on $R$.

(a) *No solution of* (4) *which is not identically zero vanishes more than once.*

(b) *There shall exist a positive function $M(s)$ defined for all positive values of $s$ and a positive constant $s^*$, both independent of $g$, such that $M(s)$ becomes positively infinite with $s$, and such that any solution of* (4) *for which* (5) *holds shall satisfy the relation*

$$|w(-s_1)| + |w(s_2)| \geqq M(s) \tag{6}$$

*provided*

$$s_1 > s > s^*$$
$$s_2 > s > s^*.$$

It is understood that (6) is independent of the geodesic $g$ upon which the solution $w(x)$ is based and of the point on $g$ from which $x = s$ is measured. We show that uniform instability implies the type hypothesis and hence regional transitivity on $R^*$. We thus have the following theorem.

THEOREM 3. *If the geodesics on $R^*$ are uniformly unstable the set of completely transitive geodesics has the power of the continuum.*

By a *geodesic ray* on $R$ we mean a geodesic which starts from a point $P$ of $R$ and is continued indefinitely on $R$ in one sense from $P$. In general there are two boundary geodesic rays of class $A$ issuing from $P$ of the "type" of any $H$-ray issuing from $P$. We show that the boundary geodesic rays of a given type are unique whenever the type hypothesis holds. The elements on a geodesic ray possess congruent elements on $S_0$, and these define a motion in the phase space $M$. If the closure of this motion on $M$ is identical with $M$ we term the geodesic ray completely transitive. We state the following theorem.

THEOREM 4. *If the unending geodesics on $R^*$ are uniformly unstable the set of completely transitive geodesic rays issuing from a fixed point $P$ on $R$ has the power of the continuum.*

*The Transitivity Function.*—We now take up the question of the deferring of the degree of transitivity. Let $g$ be a transitive geodesic ray issuing from a point $P$. If $e$ is any positive number less than a suitably chosen positive constant $p$ we show that there exists a least positive number

$$L_g(e) \qquad 0 < e < p$$

such that a segment of $g$ of length $L_g(e)$ with initial point at $P$ defines a curve $\gamma$ in the phase space $M$ at a distance at most $e$ from each element of $M$. We term $L_g(e)$ the transitivity function defined by $g$.

Our theorem on the deferring of transitivity is the following:

THEOREM 5. *Let $P$ be any arbitrary point on $S$ and $\psi(e)$ an arbitrary positive function of $e$ defined on the interval $0 < e < p$ and such that $\psi(e)$ becomes infinite as $e$ tends to zero. There then exists a completely transitive geodesic ray issuing from $P$ for which the corresponding transitivity function exceeds $\psi(e)$ at an infinite sequence of positive values of $e$ tending to zero, provided the geodesics on $R^*$ are uniformly unstable.*

*Relative Transitivity.*—We now drop the type hypothesis and the hypothesis of uniform instability. We say that a geodesic ray $g$ on $R^*$ is transitive relative to a set $H$ of geodesics on $R^*$ if the geodesics of $H$ are limit geodesics of $g$. We have proved the following theorem.

THEOREM 6. *If $P$ is an arbitrary point on $S$, the geodesic rays issuing from $P$ which are transitive relative to the set of all boundary geodesics of class $A$ have the power of the continuum.*

If the geodesics of the system possess uniform instability the boundary geodesics of class $A$ are the only geodesics and Theorem 6 implies Theorem 4. This matter of relative transitivity is obviously related to the question of the existence of invariant sets with intermediate measures in the phase space $M$.

[1] Birkhoff and Koopman, "Recent Contributions to the Ergodic Theory," *Proc. Nat. Acad. Sci.*, **18**, 279 (1932). References are here given to Hopf, Koopman, v. Neumann and P. Smith.

[2] Morse, Marston, "A Fundamental Class of Geodesics on Any Closed Surface of Genus Greater Than One," *Trans. Amer. Math. Soc.*, **26**, 25–61 (1924).

[3] Hedlund, Arnold, "On the Measure of the Non-Special Geodesics on a Surface of Constant Negative Curvature," *Proc. Nat. Acad. Sci.*, **19**, 345 (1933).

[4] Birkhoff, "Dynamical Systems Colloquium Lectures." A brief proof of regional transitivity in the case of a special closed surface of genus 2 of non-constant negative curvature is given here.

[5] Birkhoff, "Probability and Physical Systems," *Bull. Amer. Math. Soc.*, **34** (1932). The author's "Transitivity Functions" and Birkhoff's "Ergodic Function" are different but related functions.

---

## *ON DISTRIBUTION FUNCTIONS AND THEIR LAPLACE-FOURIER TRANSFORMS*

BY E. K. HAVILAND

DEPARTMENT OF MATHEMATICS, THE JOHNS HOPKINS UNIVERSITY

Communicated December 5, 1933

In the May issue of these PROCEEDINGS,[1] the author established the existence of a distribution function for functions having certain time averages, the proof being based on theorems on Radon integrals. In the May issue of the *Mathematische Annalen*, there appeared an article by Bochner,[2] of which a portion was devoted to similar theorems on Radon integrals. Bochner's considerations were based on additive interval functions rather than on the more general absolutely additive set functions.[3] Our present paper treats certain properties of distribution functions and their Laplace-Fourier transforms. Bochner shows that the product of two such transforms is a characteristic function,[4] but establishes the existence of the Faltung[5] and proves our Theorem V of the present paper only under severe restrictions, which are, however, sufficient for his purpose.

We first consider the Radon integral[6] of certain discontinuous functions. Our discussion, while limited to two dimensions for convenience, admits

of immediate extension to $n$ dimensions. Let $\varphi$ be an absolutely additive set function defined for a set[7] $T$ of sets in the $(x, y)$-plane. Let $f(x, y)$ be a bounded single-valued point function defined on the bounded set $E_0$ belonging to $T$. Let $\Pi$ denote a division of $E_0$ into subsets $E_1, \ldots E_n$, all belonging to $T$, and $D(\Pi)$ the greatest diameter of any of the sets $E_1, \ldots E_n$. Finally, let $U_i$ denote the upper limit and $L_i$ the lower limit of $f$ in $E_i$. Then if

$$\lim \sum_{k=1}^{m} U_k \varphi(E_k) \quad \text{and} \quad \lim \sum_{k=1}^{m} L_k \varphi(E_k), \quad \text{as } D(\Pi) \longrightarrow 0,$$

exist and are equal, the Radon integral of $f$ over $E_0$ is said to exist. A sufficient condition for the existence of the integral is given by

THEOREM I.[8] Let $f(x, y)$ be continuous in $\bar{J}$: $(-M \leqq x \leqq M; -M \leqq y \leqq M)$ save on the at most denumerable set of exceptional lines $x = \xi_i$, $y = \eta_j$, $i, j = 1, 2, \ldots$. Furthermore, to an arbitrarily preassigned $\epsilon < 0$ there shall correspond only a finite number $\mu = \mu(\epsilon)$, $\nu = \nu(\epsilon)$ of exceptional lines $\xi_i$, $\eta_j$, respectively, which contain all points $(x', y')$ to which there corresponds no $\delta > 0$ such that $| f(x', y') - f(x, y) | < \epsilon$, provided $| x' - x | < \delta$ and $| y' - y | < \delta$. Let $\varphi(E)$ be an absolutely additive set function defined in $J$. Then if $E_0$ be a closed point set in $J$ belonging to the domain of definition of $\varphi$, a necessary and sufficient condition that the Radon integral $\int\int_{E0} f(x, y) d_{xy}\varphi(E)$ exist is that the singular lines[1] of $\varphi$ do not coincide with the exceptional lines of $f$.

PROOF: If $u = u(\Pi) = \sum_{k=1}^{m} U_k \varphi(E_k)$ and $l = l(\Pi) = \sum_{k=1}^{m} L_k \varphi(E_k)$ and if $\Pi$, $\Pi'$ are two divisions of $E_0$ and $\Pi\Pi'$ the division consecutive to them, then $l(\Pi) \leqq l(\Pi\Pi') \leqq u(\Pi\Pi') \leqq u(\Pi')$, so that the upper limit of the $l(\Pi)$ is not greater than the lower limit of the $u(\Pi)$. Therefore the integral exists if and only if $| u(\Pi) - l(\Pi) | < \epsilon$, provided $D(\Pi) < \delta$, where $\epsilon > 0$ is arbitrarily small and $\delta = \delta(\epsilon)$. That the condition of the theorem is necessary follows at once. Its sufficiency may be proved as follows:

For a preassigned $\epsilon_1$, the $\mu(\epsilon_1)$ exceptional lines $x = \xi_i$ and the $\nu(\epsilon_1)$ exceptional lines $y = \eta_j$ can be imbedded in strips

$$X_i : (\xi_i - 2\zeta < x < \xi_i + 2\zeta), \qquad i = 1, 2, \ldots \mu$$
$$Y_j : (\eta_j - 2\zeta < y < \eta_j + 2\zeta), \qquad i = 1, 2, \ldots \nu$$

whose widths $4\zeta$ are arbitrarily small. Then by taking the subdivision sufficiently fine, $E_0$ may for this subdivision be divided into two parts $E_0'$ and $E_0''$ where $E_0'$ includes all points of $E_0$ in $W_i : (\xi_i - \zeta \leqq x \leqq \xi_i + \zeta)$, $i = 1, 2, \ldots \mu$ and $Z_j : (\eta_j - \zeta \leqq y \leqq \eta_j + \zeta)$, $j = 1, 2, \ldots \nu$, but is

included in $\sum_{i=1}^{\mu} X_i + \sum_{j=1}^{\nu} Y_j$, and $E_0'' = E_0 - E_0'$. Let $E_k', E_k''$ denote those $E_k$ composing $E_0', E_0''$, respectively. Then $u(\Pi) = u'(\Pi) + u''(\Pi)$ and $l(\Pi) = l'(\Pi) + l''(\Pi)$, where we separate the contributions to the sums made by $E_0'$ and by $E_0''$. Then $| u(\Pi) - l(\Pi) | \leqq | u'(\Pi) - l'(\Pi) | + | u''(\Pi) - l''(\Pi)|$. Now $| u''(\Pi) - l''(\Pi) | \leqq \epsilon_1 \int\int_{E_0} | d\varphi(E) | < {}^1/_2\epsilon$ by suitable choice of $\epsilon_1$ provided $D(\Pi) < \delta = \delta(\epsilon_1, \zeta)$, for as $E_0''$ lies in the smallest closed region containing $J - J\left(\sum_{i=1}^{\mu} W_i + \sum_{j=1}^{\nu} Z_j\right)$, one can obtain a subdivision $\Pi$ of $E_0$ so fine that $| U_k - L_k | < \epsilon_1$ for every $E_k''$.

Again, $| u'(\Pi) - l'(\Pi) | \leqq 2B\epsilon_2$, where $B = fin\ sup\ | f(x, y) |$ in $E_0$ and $\epsilon_2 = \sum_{i=1}^{\mu} \varphi(X_i') + \sum_{j=1}^{\nu} \varphi(Y_j')$ and $X_i' = X_iJ$, $Y_j' = Y_jJ$. Since the lines $x = \xi_i$, $y = \eta_j$ are exceptional lines of $f(x, y)$, they are by hypothesis non-singular lines of $\varphi(E)$, so that $\epsilon_2$ may be made $< {}^1/_2\epsilon$ by a suitable choice of $\zeta$.[9] It follows that $| u(\Pi) - l(\Pi) | < \epsilon$ when $D(\Pi) < \min(\zeta, \delta)$, where $\delta = \delta(\epsilon, \zeta)$. Hence as $D(\Pi) \longrightarrow 0$, both $\lim \sum_{k=1}^{m} U_k\varphi(E_k)$ and $\lim \sum_{k=1}^{m} L_k\varphi(E_k)$ exist and are equal, q.e.d.

We proceed to apply Theorem I to a special case, making use of the

DEFINITION. The monotone absolutely additive set function $\varphi(E)$ is said to be a distribution function if $0 \leqq \varphi(E) \leqq 1$ and $\varphi(S) = \int\int_S d_{xy}\varphi(E) = 1$, where $S$ denotes the whole plane.

If in addition $\varphi(E) = 0$ for all sets $E$ outside a fixed rectangle $J$, we shall say that $\varphi(E)$ is damped. In this case, the integral over $S$ is equivalent to the integral over $J$. In what follows, we shall confine ourselves for simplicity to damped distribution functions. This restriction is in no wise essential and will be avoided in a more detailed paper to be published shortly. We may now state

THEOREM II. If $\varphi_1(E)$ and $\varphi_2(E)$ be damped distribution functions, then there exists an everywhere dense[9] sequence of rectangles $R_n$ with sides parallel to the axes such that the Radon integral[10]

$$\Phi(Q) = \int\int_S \varphi_1(Q - P_{xy}) d_{xy}\varphi_2(E)$$

exists for $Q = R_n$, $n = 1, 2, \ldots$.

PROOF: In the first place, the exceptional lines of the point function $f(x, y) = \varphi_1(Q - P_{xy})$, where $Q$ is fixed, are those lines $x = x_i$, $y = y_j$ such that at least one side of $Q - P_{xy}$, $Q$ being a rectangle with sides parallel to the axes, lies on one of the singular lines $x = \xi_i'$, $y = \eta_j'$ of $\varphi_1(E)$. As $Q$ has precisely two sides parallel to each axis and as the singular lines of $\varphi_1(E)$ are at most denumerable,[11] it follows that so also are the exceptional lines of the point function $f(x, y) = \varphi_1(Q - P_{xy})$. Moreover, as

$\varphi_1(E)$ is bounded and monotone, the points at which its saltus (i.e., the two-dimensional saltus of the corresponding point function) exceeds $\epsilon$ are confined to a finite number $n = n(\epsilon)$ of its singular lines.[11] Accordingly, the number of lines containing at least one point at which the saltus of $\varphi_1(Q - P_{xy})$ exceeds $\epsilon$ is finite ($= 2n(\epsilon)$) for any given rectangle $Q$.

If the singular lines of $\varphi_2(E)$ be denoted by $x = \xi_i$, $y = \eta_j$, the number of sides of all rectangles $Q$ for which $x_i = \xi_p$, $y_j = \eta_q$, $i, j, p, q = 1, 2, 3, \ldots$ is clearly at most denumerable. Consequently, there is an everywhere dense set of rectangles $R_n$ such that the exceptional lines of $f(x, y) = \varphi_1(Q - P_{xy})$ are distinct from the singular lines of $\varphi_2(E)$. Applying Theorem I we are then led to the result stated in Theorem II.

From the definition of $\Phi(R_n)$ and the properties of $\varphi_1$, $\varphi_2$ it follows at once that $\Phi(R_n)$ is a monotone function of $R_n$. As $\varphi_1(R_n - P_{xy})$ is zero for all $R_n$ so distant, say outside a rectangle $J_1$, that $R_n - P_{xy}$ lies outside $J$ for all $P_{xy}$ in $J$, it follows that $\Phi(R_n) = 0$ for all $R_n$ outside $J_1$, while if $R_n$ is taken so large that $R_n - P_{xy}$ includes $J$ for all $P_{xy}$ in $J$, it is easy to show that $\Phi(R_n) = 1$ for all $R_n$ enclosing a certain rectangle $J_2$.

Now among the rectangles for which the integral $\Phi$ is defined is a sequence of rectangles $R_s : (-K \leqq x < x_s;\ -K \leqq y < y_s)$ such that the points $(x_s, y_s)$ are everywhere dense in $J_3 : (-K \leqq x < K;\ -K \leqq y < K)$, a rectangle containing $J_1$ and $J_2$ in its interior. Moreover, the points $(x_s, y_s)$ are everywhere dense on the parallels to the axes through any one of them and they form the vertices of a set of rectangles everywhere dense in $J_3$. Then to $\Phi = \Phi(R)$ we may assign a point function $F^*$ defined only at the points $(x_s, y_s)$ by the relations

$$F^*(x_s, y_s) = \Phi(R_s) \qquad \text{and} \qquad F^*(-K, y_s) = F^*(x_s, -K) = 0.$$

On the points $(x_s, y_s)$ we shall prove $F^*(x_s, y_s)$ possesses the following properties:

$(\alpha)$ $F^*(x + h, y + k) - F^*(x + h, y) - F^*(x, y + k) + F^*(x, y) \geqq 0$, where $h \geqq 0$, $k \geqq 0$ and $(x + h, y + k)$, $(x + h, y)$, $(x, y + k)$, $(x, y)$ are points of the everywhere dense sequence $\{(x_s, y_s)\}$.

$(\beta)$ $\lim_{h=0,\ k=0} F^*(x - h, y - k) = F^*(x, y)$, where $h \geqq 0$, $k \geqq 0$, and $(x - h\ y - k)$, $(x, y)$ belong to the everywhere dense sequence.

$(\gamma)$ $F^*(-K, y) = F^*(x, -K) = 0$ and $F^*(K, K) = 1$, where $(-K, y)$, $(x. -K)$ and $(K, K)$ are points of the everywhere dense sequence.

PROOF OF $(\alpha)$: As $\varphi_1$ is additive, it may readily be shown that $\Phi(R) = \Phi(R_1) + \Phi(R_2)$, where $R = R_1 + R_2$ and $R_1$, $R_2$ are both rectangles for which $\Phi$ is defined. This fact, together with the monotone character of $\Phi(R)$, leads to the correctness of $(\alpha)$.

PROOF OF $(\beta)$: It is to be shown that if $R^0 : (-K \leqq x < x^0;\ -K \leqq y < y^0)$ is an arbitrary fixed rectangle of $\{R_s\}$ and $R_s : (-K \leqq x < x_s;$

$-K \leqq y < y_s$) a variable rectangle of the sequence, then $0 \leqq \Phi(R^0) - \Phi(R_s) < \epsilon$, provided $0 \leqq x^0 - x_s < \delta^0$, $0 \leqq y^0 - y_s < \delta^0$, where $\delta^0 = \delta^0(\epsilon, x^0, y^0)$. The proof follows that of Theorem I. Let $W$ be a closed rectangle so large that it contains in its interior the points $(u, v)$, where $u = x^0 - x$, $v = y^0 - y$ for all $(x, y)$ in $J$. With $\varphi_1$ there is associated[12] a point function $F_1(u, v)$ such that $| F_1(u, v) - F_1(u', v') | < {}^1/_2\epsilon$, provided $| u - u' | < \delta = \delta({}^1/_2\epsilon, u, v)$, $| v - v' | < \delta = \delta({}^1/_2\epsilon, u, v)$ for all points of $W$ except at most the points of $\mu = \mu(\epsilon)$ lines $u = u_i$ and $\nu = \nu(\epsilon)$ lines $v = v_j$. To these lines there correspond lines $x = \xi_i$, $y = \eta_j$ which by the choice of $R^0$ are non-singular lines of $\varphi_2(E)$ so that they may be imbedded in a finite number of open strips $X_i$, $Y_j$ so narrow that $0 \leqq \sum_{i=1}^{\mu} \varphi_2(X_i) + \sum_{j=1}^{\nu} \varphi_2(Y_j) < \epsilon/4$. To these strips correspond strips $U_i$, $V_j$ containing the lines $u = u_i$, $v = v_j$, respectively. In the closed region $W \cdot [S - (\sum_{i=1}^{\mu} U_i + \sum_{j=1}^{\nu} V_j)] = W_1$, $\delta({}^1/_2\epsilon, u, v) \geqq \delta'({}^1/_2\epsilon) > 0$, where $\delta'({}^1/_2\epsilon)$ is independent of the point $(u, v)$ there, so that $| \varphi_1(R^0 - P_{xy}) - \varphi_1(R_s - P_{xy}) | < {}^1/_2\epsilon$ for all $(x, y)$ such that $(x^0 - x, y^0 - y)$ lies in $W_1$, i.e., for all $(x, y)$ in $J \cdot [S - (\sum_{i=1}^{\mu} X_i + \sum_{j=1}^{\nu} Y_j)] = J_4$, and for $R_s$ such that $| x^0 - x_s | < \delta'$, $| y^0 - y_s | < \delta'$, where $\delta' = \delta'({}^1/_2\epsilon)$ is independent of $(x, y)$ in $J_4$. Consequently,

$$| \int\int_{J_4} [\varphi_1(R^0 - P_{xy}) - \varphi_1(R_s - P_{xy})] d_{xy}\varphi_2(E) | < {}^1/_2\epsilon$$

if $| x^0 - x_s | < \delta'$, $| y^0 - y_s | < \delta'$. Moreover,

$$| \int\int_{J - J_4} [\varphi_1(R^0 - P_{xy}) - \varphi_1(R_s - P_{xy})] d_{xy}\varphi_2(E) | \leqq 2\left[\sum_{i=1}^{\mu} \varphi_2(X_i) + \sum_{j=1}^{\nu} \varphi_2(Y_j)\right] < {}^1/_2\epsilon,$$

so that in particular as integration over the whole space $S$ is effectively integration over $J$,

$$0 \leqq \int\int_S [\varphi_1(R^0 - P_{xy}) - \varphi_1(R_s - P_{xy})] d_{xy}\varphi_2(E) < \epsilon,$$

if $0 \leqq x^0 - x_s < \delta^0 = \delta'$, $0 \leqq y^0 - y_s < \delta^0 = \delta'$,

i.e., $F^*(x_s, y_s)$ possesses property $(\beta)$.[13]

PROOF OF $(\gamma)$: That $F^*(x_s, y_s)$ possesses property $(\gamma)$ follows from its definition and that of $J_3$.

Now it is known[14] that if a function $F^*$, defined for a set of points everywhere dense in $J_3$ and which are also such that they are everywhere dense on the parallels to the axes through any one of them and that they include the corners of an everywhere dense set of rectangles, possesses

properties $(\alpha)$, $(\beta)$, $(\gamma)$, there exists at least one function $F$ defined throughout $J_3$ and such that $F$ fulfills conditions $(\alpha)$, $(\beta)$, $(\gamma)$ there and $F = F^*$ on all points of the everywhere dense set. Furthermore, $F$ is determined by $F^*$ up to the singular lines of $F$. To this function $F(x, y)$ there corresponds[12] a monotone absolutely additive set function $\psi(E)$, defined throughout the entire plane if we set $\psi(E) = 0$ for all sets $E$ outside $J_3$, and such that $\psi(S) = 1$. We have thus proved

THEOREM IV. There exists a damped distribution function $\psi(E)$ such that $\psi(R) = \Phi(R)$ on all rectangles $R$ for which $\Phi$ is defined. Furthermore, $\psi(E)$ is uniquely determined up to its singular lines.

An important property of $\psi(E)$ is given by

THEOREM V.[15] $L(s, t; \psi) = L(s, t; \varphi_1)L(s, t; \varphi_2)$, where $L(s, t; \sigma) = \int\int_S \exp[i(sx + ty)]d_{xy}\sigma(E)$ and $s$ and $t$ are real.

PROOF: As all three integrals clearly exist, they are by their definition as Radon integrals limits of approximating sums in which the $E_k$ may be chosen as rectangles $R_k$ of the everywhere dense set for which $\Phi$ is defined. By choosing the $R_k$ sufficiently small and $A = \sum_{k=1}^{N} R_k$ sufficiently large that it enclose $J_3$ we may obtain

$$\left| \int\int_S \exp[i(sx + ty)]d_{xy}\psi(E) - \sum_{k=1}^{N} \exp[i(sx_k + ty_k)]\psi(R_k) \right| < {}^1/_2\epsilon \quad (1)$$

where $N = N(\epsilon)$. If in addition $A$ be chosen so large that $A - P_j = \sum_{k=1}^{N} (R_k - P_j)$ includes $J$ for all points $P_j$ in $J$, we have also for a sufficiently fine mesh

$$\sum_{k=1}^{N} \exp[i\{s(x_k - x_j) + t(y_k - y_j)\}]\varphi_1(R_k - P_j) = \int\int_S \exp[i(sx + ty)]d_{xy}\varphi_1(E) + \zeta_{1j} \quad (2)$$

where $|\zeta_{1j}| < \epsilon$ and it is to be noted that the fineness of the $R_i$ is independent of $P_j$, as is also $\epsilon$, so that $N = N(\epsilon)$ is fixed. The meshes may then be so further subdivided that

$$\left| \psi(R_k) - \sum_{j=1}^{m} \varphi_1(R_\kappa - P_j)\varphi_2(E_j) \right| < \epsilon/2N, \quad k = 1, 2, \ldots N = N(\epsilon)$$

where $m = m(\epsilon)$ and $P_j$ is a point of $E_j$. Hence

$$\exp[i(sx_k + ty_k)]\psi(R_k) - \sum_{j=1}^{m} \exp[i(sx_k + ty_k)]\varphi_1(R_k - P_j)\varphi_2(E_j)| < \epsilon/2N.$$

On summing with respect to $k$, we have, with the help of (1), after interchanging the order of the double summation:

$$\left| \int\int_S \exp[i(sx + ty)]d_{xy}\psi(E) - \sum_{j=1}^{m}\sum_{k=1}^{N} \exp[i\{s(x_k - x_j) + t(y_k - y_j)\}] \right.$$
$$\left. \varphi_1(R_k - P_j) \cdot \exp[i(sx_j + ty_j)]\varphi_2(E_j) \right| < \epsilon. \quad (3)$$

Let $m$ be so large and $T = \sum_{j=1}^{m} E_j$ so large (if this is not already the case) that

$$\sum_{j=1}^{m} \exp[i(sx_j + ty_j)]\varphi_2(E_j) = \int\int_S \exp[i(sx + ty)]d_{xy}\varphi_2(E) + \zeta_2, \quad (4)$$
$|\zeta_2| < \epsilon$.

Then substituting (2) and (4) in (3), we obtain

$$\left| \int\int_S \exp[i(sx + ty)]d_{xy}\psi(E) \right.$$
$$- \int\int_S \exp[i(sx + ty)]d_{xy}\varphi_1(E) \cdot \int\int_S \exp[i(sx + ty)]d_{xy}\varphi_2(E)$$
$$\left. - \zeta_2 \int\int_S \exp[i(sx + ty)]d_{xy}\varphi_1(E) - \sum_{j=1}^{m} \zeta_{1j} \exp[i(sx_j + ty_j)]\varphi_2(E_j) \right| < \epsilon$$

so using the appraisals for $\zeta_2$ and $\zeta_{1j}$ and the fact that the total variation of $\varphi_1$ and of $\varphi_2$ is 1, we find

$$\left| \int\int_S \exp[i(sx + ty)]d_{xy}\psi(E) \right.$$
$$\left. - \int\int_S \exp[i(sx + ty)]d_{xy}\varphi_1(E) \cdot \int\int_S \exp[i(sx + ty)]d_{xy}\varphi_2(E) \right| < 3\epsilon$$

q.e.d.

It may be proved directly by considering the finite sums analogous to that in (2) that $\psi(E)$ as above defined is not affected by interchanging $\varphi_1$ and $\varphi_2$ in the definition of $\Phi(Q)$. This proof is not given here as the result is also a consequence of Theorem V and the fact that the distribution functions are determined up to their singular lines by their Laplace-Fourier transforms.[16]

[1] E. K. Haviland, these PROCEEDINGS, **19**, 549–555 (1933). It is not required that the function be almost-periodic but only that it be an $H$-function in the sense of Wintner (these PROCEEDINGS, **18**, 248–251 (1932)). This $H$-condition seems somewhat special, but in reality Birkhoff's general ergodic theorem shows that for the solutions of any dynamical system having a finite definite integral invariant, the $H$-condition is always fulfilled save at most for solutions having the geometrical probability zero (G. D. Birkhoff, these PROCEEDINGS, **17**, 650–660 (1931)).

[2] S. Bochner, *Math. Annalen*, **108**, 378–410 (1933).

[3] This specialization is not as great as might at first appear as it can be shown that to a monotone additive interval function there corresponds a monotone absolutely additive set function uniquely up to the singular lines of the latter. This result is substantially given by Bochner, loc. cit., p. 387.

[4] S. Bochner, *Ibid.*, p. 399.

[5] For comment on this term, cf. N. Wiener, *The Fourier Integral*, p. 45 (1933).

[6] For definition, cf. J. Radon, *Wiener Sitzungsber.*, **122**, 1322–1324 (1913).

[7] Cf. J. Radon, *Ibid.*, p. 1299.

[8] For a treatment of one-dimensional Stieltjes integrals, cf. R. Schmidt, *Math. Zeitschr.*, **22**, 123–125 (1925) and in a general form, E. W. Hobson, *Functions of a Real*

*Variable*, 2nd Ed., Vol. I, pp. 506–512 (1921). Extensions of our Theorem I in the manner of Hobson are easily possible but unnecessary for our present purposes.

[9] E. K. Haviland, loc. cit., p. 550.

[10] $P_{xy}$ is the point with the coördinates $(x, y)$. $Q - P_{xy}$ is to be taken in the sense of vector addition. Cf., e.g., H. Bohr and B. Jessen, *D. Kgl. Danske Vidensk. Selsk. Skrifter*, Ser. 8, 12, 332 (1929).

[11] Cf. J. Radon, *Wiener Sitzungsber.*, **128**, 1093 (1919).

[12] Cf. Radon, loc. cit., pp. 1304–1305.

[13] The foregoing proof contains the stronger result that $(x^0, y^0)$ is a continuity point of $F^*$ and hence of $F$, and this is true of every $(x^0, y^0)$ such that the exceptional lines of $\varphi_1(R^0 - P_{xy})$ are distinct from the singular lines of $\varphi_2(E)$. Thus the singular lines of $F$ are at most those whose form is $x = \xi_i + \xi'_p$, $y = \eta_j + \eta'_q$.

[14] Cf. J. Radon, loc. cit., pp. 1338–1339. I have perceived since writing Note 14 to my above-mentioned article that Radon does not need to show that his function $F^*$ defined on the everywhere dense set of points possesses property $(\beta)$ there. However, when such is the case it is possible at once to infer that $F = F^*$ on all points for which the latter is defined.

[15] Cf., e.g., v. Mises, *Wahrscheinlichkeitsrechnung u. ihre Anwendung*, p. 219 (1931). This is the so-called independence relation often used in a formal way. We are not aware of a published general proof of our Theorem V even in the one-dimensional case.

[16] Cf., e.g., S. Bochner, loc. cit., p. 402.

---

## *ON THE ASYMPTOTIC FORMULAE OF RIEMANN AND OF LAPLACE*

BY AUREL WINTNER

THE JOHNS HOPKINS UNIVERSITY, DEPARTMENT OF MATHEMATICS

Communicated December 5, 1933

*The Riemann Case.*[1]—Let $f(x)$ denote a function having a continuous second derivative in the finite range $0 \leqq x \leqq a$, and let $t$ be a positive parameter. Perron[2] has proved that

$$\int_0^a f(\lambda) \exp(\pm itx^2)dx = (\pi/2^3)^{-1/2}(1 \pm i)f(0)t^{-1/2} + O(t^{-1}) \quad (1)$$

as $t \longrightarrow +\infty$ and applied this formula for the determination of the asymptotic behavior of functions[1]

$$c(t) = \int_\alpha^\beta \phi(\lambda) \exp[it\omega(\lambda)]d\lambda \quad (2)$$

where $\omega(\lambda)$ is real-valued and not a constant and $\omega(\lambda)$, $\phi(\lambda)$ are, for instance, regular-analytic in the interval $\alpha \leqq \lambda \leqq \beta$. Since the Bessel functions $J_p$ are integrals $c(t)$ of the type (2), there follows in particular from (1) the classical relation $J_p(t) \sim (2/\pi)^{1/2} \cos(t - \pi p/2 - \pi/4)t^{-1/2}$.

The reduction of the asymptotic behavior of (2) to the normalized formula (1) is, however, possible only if the derivative $\omega'(\lambda)$ of the frequency nowhere possesses a multiple zero, as is the case for $J_\phi(t)$. The vicinity of a point $\lambda = \lambda_0$ at which $\omega'(\lambda_0) \neq 0$ yields to (2) a contribution which is but of the order of the remainder $O(t^{-1})$ in (1), inasmuch as

$$\int_0^a f(x) \exp(\pm itx)dx = O(t^{-1}) \tag{3}$$

in virtue of the second mean-value theorem. In the first part of the present note Perron's proof for (1) will be slightly modified and simultaneously extended in such a way that the treatment of (2) is possible even if $\omega'(\lambda)$ does possess a multiple zero. In order to obtain this extension one clearly has to generalize (1) for the case where $\exp(\pm itx^2)$ is replaced by $\exp(\pm itx^n)$, $n \geqq 2$. The result will be seen not to be the same as that for $n = 2$, although just as simple, viz.,

$$\int_0^a f(x) \exp(\pm itx^n)dx = \exp(\pm \pi i/2n)\, \Gamma(1 + n^{-1})f(0)t^{-1/n} + O(t^{-2/n}) \tag{4}$$

where $t \longrightarrow +\infty$, $n \geqq 2$. For $n = 2$ one obtains precisely (1) inasmuch as $2\Gamma(3/2) = \Gamma(1/2) = \pi^{1/2}$. For $n = 1$, of course (4) is false as illustrated by examples as $f(x) = 1$ or $f(x) = x$. Since $t^{-2/n} = O(t^{-1})$ for every $n \geqq 2$, it is clear from (4) and (3) that the contribution of the vicinity of a point $\lambda = \lambda_0$ to the integral (2) always is absorbed by the $O$-term in (4) if $\omega'(\lambda_0) \neq 0$. In the proof of (4) it will not be supposed that $f(x)$ is analytic or that $n(\geqq 2)$ is an integer.

Suppose that $f(x)$ possesses a continuous first derivative in the whole interval $0 \leqq x \leqq a$ and a continuous second derivative in an arbitrarily narrow interval $0 \leqq x \leqq b$ where $0 < b \leqq a$. On placing

$$xg(x) = f(x) - f(0) - xf'(0) \quad \text{if} \quad 0 < x \leqq a \tag{5}$$

and $g(x) = 0$ if $x = 0$, it follows by a repeated application of the mean-value theorem of the differential calculus that $g(x)$ possesses a continuous first derivative not only in the range $0 < x \leqq a$ but in the closed interval $0 \leqq x \leqq a$ as well. Furthermore

$$\int_0^a xg(x) \exp(itx^n)dx = O(t^{-2/n}). \tag{6}$$

For on applying partial integration the integral (6) may be written in the form

$$g(a) \int_0^a x \exp(itx^n)dx - \int_0^a g'(x) \int_0^x y \exp(ity^n)dy\, dx$$

or

$$t^{-2/n}[g(a)H(at^{1/n}) - \int_0^a g'(x)H(xt^{1/n})dx] \tag{6a}$$

where $H(t) = \int^t y \exp(iy^n)dy$. Consequently, on writing $y$ instead of $y^n$,

$$H(t^n) = \int_0^t y^{-1+2/n} \exp(iy)dy/n. \tag{6b}$$

Hence $H$ is a bounded function of $t$ for every fixed $n \geqq 2$. For if $n > 2$ so that the exponent of $y$ in (6b) is $< 0$, then on applying the second mean-value theorem there follows the existence of the improper integral $H(+\infty)$. If, on the other hand, $n = 2$ then $|H| \leqq 1$ is obvious from (6b). Since $H$ remains bounded as $t \longrightarrow +\infty$ and since $g'(x)$ is continuous and therefore bounded in the finite range $0 \leqq x \leqq a$, the expression (6a) clearly is $= O(t^{-2/n})$ as stated under (6).

Furthermore,

$$\int_0^a x \exp(itx^n)dx = O(t^{-2/n}). \tag{7}$$

For on placing $tx^n = y$, the integral (7) takes the form $t^{-2/n} H(a^n t^n)$ where $H$ is, as we saw, a bounded function.

Finally,

$$\int_a^{+\infty} \exp(itx^n)dx = O(t^{-2/n}). \tag{8}$$

In order to prove (8) where $a > 0$ and $n \geqq 2$, we first notice that the integral

$$L_n(t) = \int_t^{+\infty} y^{-1+1/n} \exp(iy)dy$$

where $n > 1$ exists and is $= O(t^{-1+1/n})$ in virtue of the second mean-value theorem applied to a finite interval. Now it follows from $L_n(t) = O(t^{-1+1/n})$ that the integral (8) which may be written in the form $t^{-1/n} L_n(a^n t)/n$ is $= t^{-1/n} O(t^{-1+1/n}) = O(t^{-1})$ if $n > 1$, hence *a fortiori* $= O(t^{-2/n})$ if $n \geqq 2$.

On substituting (5) and (7) into (6) and combining the result with (8) one obtains

$$\int_0^a f(x) \exp(itx^n)dx = f(0) \int_0^{+\infty} \exp(itx^n)dx + O(t^{-2/n}) \tag{9}$$

inasmuch as the complementary integral (8) is absorbed by the $O$-term in (9). Since $\Gamma(1+z) = z\Gamma(z)$ implies

$$\Gamma(1+n^{-1}) = n^{-1} \int_0^{+\infty} x^{-1+1/n} \exp(-x)dx = \int_0^{+\infty} \exp(-r^n)dr$$

where $r^n = x$, there follows from Cauchy's theorem with the help of a standard appraisal the identity[3]

$$\int_0^{+\infty} \exp(ix^n)dx = \Gamma(1 + n^{-1}) \exp(\pi i/2n).$$

Hence the asymptotic formula (4) is a consequence of (9) inasmuch as both members of (4) go over into conjugated complex values if one replaces $i$ by $-i$. Besides (4) is an asymptotic formula only if $f(0) \neq 0$ and otherwise a mere appraisal.

*The Laplace Case.*[4]—Let $\phi(\lambda)$ and $\psi(\lambda)$ denote two regular-analytic functions in the interval $\alpha \leqq \lambda \leqq \beta$, and let $\psi(\lambda)$ be real-valued and nowhere negative in this interval. The question regarding the asymptotic behavior of the function

$$C(t) = \int_\alpha^\beta \phi(\lambda)[\psi(\lambda)]^t d\lambda \quad (t > 0, \quad \psi^t \geqq 0) \tag{2a}$$

for large values of $t > 0$ is known to be reducible to the corresponding question regarding the normalized functions

$$C_n(t) = \int_0^a f(x) \exp(-tx^n)dx \qquad (n = 1, 2, \ldots). \tag{2b}$$

It has been shown by Perron[5] that the Laplace formula, which is analogous to (1), is for all these "functions of large numbers" (2a), (2b) but the first term of an infinite asymptotic development as illustrated by the divergent Stirling series in the theory of the $\Gamma$-function. Another proof has been given by Haar.[6] The apparatus applied by Perron and Haar belongs to the theory of complex functions. It has been pointed out by the present author[7] that the asymptotic development in question may be obtained solely by methods of real analysis. This proof consisted of successive partial integrations and used an existence theorem of Borel concerning non-analytic functions with derivatives of arbitrarily high order. The application of this Borel theorem has been then eliminated by Ikehara and by Wiener[8] in the course of Wiener's Tauberian researches. The following way, suggested by the above treatment of Riemann's case, is still shorter and has the obvious advantage of being independent of a general theory.

First, if $n > 0$, $m > 0$, $a > 0$ and

$$G(x) = \int_0^x y^{m-1} \exp(-y^n)dy \tag{10}$$

then

$$\int_0^a |G(as) - G(xs)| dx = O(s^{-1}), \quad 0 < s \longrightarrow +\infty. \tag{11}$$

For the integral (11) may be written in virtue of $G(xs) \leqq G(as)$, $0 \leqq sx \leqq sa$, in the form

$$\int_0^a dx \int_{xs}^{as} y^{m-1} \exp(-y^n) dy = s^{-1} \int_0^{as} dx \int_x^{as} y^{m-1} \exp(-y^n) dy$$

where the factor of $s^{-1}$ remains bounded when $s \longrightarrow +\infty$, inasmuch as the improper integral

$$\int_0^{+\infty} K(x)dx \text{ where } K(x) = \int_x^{+\infty} y^{m-1} \exp(-y^n) dy$$

clearly exists.

Let $f(x)$ be a function posessing an $m$-th continuous derivative ($m > 0$) in the interval $0 \leqq x \leqq a$. On placing

$$x^{m-1}g(x) = f(x) - \sum_{k=1}^{m} x^{k-1} f^{(k-1)}(0)/\Gamma(k) \text{ if } 0 < x \leqq a \tag{12}$$

and $g(x) = 0$ if $x = 0$, there follows from the mean-value theorem of the differential calculus the existence and the continuity of the first derivative of $g(x)$ not only in the range $0 < x \leqq a$ but in the closed interval $0 \leqq x \leqq a$ as well. Furthermore,

$$\int_0^a x^{m-1} g(x) \exp(-tx^n) dx = O(t^{-(m+1)/n}) \quad (n > 0) \tag{13}$$

as $t \longrightarrow +\infty$. For the integral (13) takes by partial integration the form

$$g(a) \int_0^a y^{m-1} \exp(-ty^n) dy - \int_0^a g'(x) \int_0^x y^{m-1} \exp(-ty^n) dy\, dx,$$

or, on using the notation (10) and the definition $g(0) = 0$,

$$t^{-m/n} g(a) G(at^{1/n}) - t^{-m/n} \int_0^a g'(x) G(xt^{1/n}) dx$$

$$= t^{-m/n} \int_0^a g'(x) [G(at^{1/n}) - G(xt^{1/n})] dx.$$

Hence (13) is a consequence of (11) inasmuch as $|g'(x)|$ is continuous and therefore bounded in the finite interval $0 \leqq x \leqq a$.

Furthermore, if $a > 0$, $k > 0$, $n > 0$, $m > 0$ then

$$\int_a^{+\infty} x^{k-1} \exp(-tx^n) dx = O(t^{-(m+1)/n}). \tag{14}$$

For let $M = M(t)$ denote the maximum of $x^{k-1} \exp(-x^{n/2})$ in the infinite range $T \leqq x < +\infty$ where $T = at^{1/n}$ so that

$$M(t) = O(t^{-p}) \tag{15}$$

for arbitrarily large values of the exponent $p$. The integral (14) is

$$= t^{-k/n} \int_T^{+\infty} x^{k-1} \exp(-x^{n/2}) \exp(-x^n + x^{n/2}) dx$$

$$< t^{-k/n} \int_0^{+\infty} M(t) \exp(-x^n + x^{n/2}) dx = \text{const. } t^{-k/n} M(t).$$

Hence (14) is obvious inasmuch as the exponent $p$ in (15) is arbitrary and may therefore be chosen $= (m + 1 - k)/n$.

On substituting (12) into (13) and combining the result with (14) one obtains

$$\int_0^a f(x) \exp(-tx^n) dx - \sum_{k=1}^{m} \int_0^{+\infty} x^{k-1} \exp(-tx^n) dx f^{(k-1)}(0)/\Gamma(k) = O(t^{-(m+1)/n})$$

inasmuch as the complementary integrals (14) are absorbed by the $O$-term. Now on placing $tx^n = y$, the factor of $f^{(k-1)}(0)/\Gamma(k)$ in the last formula proves to be $= t^{-k/n}\Gamma(k/n)/n$. Hence

$$\int_0^a f(x) \exp(-tx^n) dx = \sum_{k=1}^{m} t^{-k/n} f^{(k-1)}(0) \Gamma(k/n)/n\Gamma(k) + O(t^{-(m+1)/n})$$

if $f(x)$ possesses an $m$-th continuous derivative. Consequently

$$\int_0^a f(x) \exp(-tx^n) dx \sim \sum_{k=1}^{\infty} t^{-k/n} f^{(k-1)}(0) \Gamma(k/n)/n\Gamma(k)$$

if all derivatives of $f(x)$ exist, e.g., if $f(x)$ is regular-analytic. This is the desired asymptotic development. It may be noted that $n(> 0)$ need not be an integer. Furthermore, the derivatives of order $k > 1$ need not exist without a small neighborhood $0 \leqq x \leqq b$ of $x = 0$ where $b < a$.

[1] Cf. B. Riemann, Gesammelte Werke, 246–248 (1876) (and, on the other hand, 400–406). Cf. also G. H. Hardy, *Quart. J. Math.*, **44**, 1–40 (1913), and 242–263 where further references also are given.

[2] O. Perron, *Archiv Math. Phys.*, Ser. III, **22**, 329–340 (1914), more particularly Hilfssatz I, $w = \pm\pi/4$.

[3] Cf. W. F. Osgood, *Lehrbuch der Funktionentheorie*, Vol. I, 4th edition, 293–294 (1923). The proof given there for $n = 2$ holds for every $n > 1$.

[4] Cf. G. Pólya-G. Szegö, *Aufgaben und Lehrsätze*, Vol. I, 244 (1924).

[5] O. Perron, *Münchner Sitzungsberichte*, 191–220 (1917).

[6] A. Haar, *Math. Annalen*, **96**, 69–107 (1926).

[7] A. Wintner, *Math. Zeitschrift*, **28**, 416–429 (1928).

[8] N. Wiener, *Ann. Math.*, Ser. II, **33**, 60–62 (1932).

# FUNCTIONS ORTHOGONAL IN THE HERMITIAN SENSE. A NEW APPLICATION OF BASIC NUMBERS

BY H. BATEMAN

DEPARTMENT OF MATHEMATICS, CALIFORNIA INSTITUTE OF TECHNOLOGY

Communicated December 12, 1933

§1. To find a particular set of functions $H_n(u)$ satisfying the Hermitian relation

$$I_{m,n} \equiv \int_{-\infty}^{\infty} e^{-1/2x^2} H_m(ix) H_n(-ix)dx = 0 \qquad (1)$$

in which the exponential factor is exp $(-x^2/2)$ as also in (14) we may put $z = e^{iax}$, where $a$ is an arbitary positive constant and assume that $H_n(ix)$ is a polynomial of the $n$th degree in $z$ with real coefficients. If $q = \exp(-a^2)$ and use is made of Euler's generalized binomial coefficient which Jensen[1] denotes by the symbol $(n, r)$ and R. Tambs Lyche[2] by the symbol $\begin{bmatrix} n \\ r \end{bmatrix}_q$, or $\begin{bmatrix} n \\ r \end{bmatrix}$, we have, with F. H. Jackson's notation $[n]$ for a basic number[3]

$$[n](1-q) = (1-q^n), \quad (n,r)[1][2]\ldots[r] = [n][n-1]\ldots [n-r+1] \quad (n,0) = 1. \qquad (2)$$

If, now

$$S_m^{(n)} = \sum_{r=0}^{n} (-)^r (n,r) q^{r(m-n)+\frac{1}{2}r(r+1)} \qquad (3)$$

we readily find that

$$S_{m+1}^{(n)} - S_m^{(n)} = (1-q^n) q^{m-n+1} S_m^{(n-1)}. \qquad (4)$$

Hence

$$S_m^{(n)} = 0 \qquad m < n$$

$$S_n^{(n)} = (1-q^n) S_{n-1}^{(n-1)} = (1-q)(1-q^2)\ldots(1-q^n). \qquad (5)$$

On account of these relations we may write

$$H_n(ix) = C_n \left\{ z^n - (n,1) q^{1/2} z^{n-1} + (n,2) q z^{n-2} \ldots + (-)^n (m,n) q^{1/2n} \right\}, \qquad (6)$$

where $C_n$ is a constant which will be chosen so that $I_{nn} = 1$. The appropriate value of $C_n$ is given by the equation

$$C_n^2 (2\pi)^{1/2} (1-q)(1-q)^2 \ldots (1-q^n) = 1. \qquad (7)$$

To expand $z^m$ in a series of functions $H_n(ix)$ we make use of the inversion formula

$$X_n = \sum_{r=0}^{n} (-)^r (n, r) q^{1/2r} Y_{n-r}$$
$$Y_m = \sum_{s=0}^{m} (m, s) q^{1/2s^2} X_{m-s} \tag{8}$$

which is easily verified with the aid of the relations (5). The relation to be established is, in fact

$$X_n = \sum_{r=0}^{n} (-)^r (n, r) q^{1/2r} \sum_{k=0}^{n-r} (n-r, n-r-k) q^{1/2(n-r-k)^2} X_k. \tag{9}$$

Changing the order of summation we have to prove that

$$0 = \sum_{r=0}^{n-k} (-)^r (n\, r)(n-r, n-r-k) q^{kr + 1/2r(r+1)} \quad k \neq n. \tag{10}$$

Now

$$(n, r)(n-r, n-r-k) = (n, k)(n-k, r) \tag{11}$$

and so the relation to be established reduces to a particular case of (5). We obtain, then, the relation

$$z^m = \sum_{r=0}^{m} (m, r) q^{1/2r^2} C_{m-r}^{-1} H_{m-r}(ix) \tag{12}$$

which may be used to transform a power series in $z$ into a series of type

$$f(x) = \sum_{n=0}^{\infty} A_n H_n(ix), \tag{13}$$

in which, under suitable conditions

$$A_n = \int_{-\infty}^{\infty} e^{-1/2x^2} H_n(-ix) f(x) dx. \tag{14}$$

Putting

$$f(x) = \sum_{m=0}^{\infty} B_m z^m \tag{15}$$

the equations of transformation are

$$B_m = \sum_{r=0}^{\infty} (-)^r (m+r, r) q^{1/2r} C_{m+r} A_{m+r}$$
$$A_n = \sum_{m=0}^{\infty} B_m \int_{-\infty}^{\infty} e^{-1/2x^2} H_n(-ix) z^m dx \tag{16}$$
$$= C_n^{-1} \sum_{s=0}^{\infty} (n+s, s) q^{1/2s^2} B_{n+s}$$

and these suggest the existence of a second inversion formula

$$X_m = \sum_{r=0}^{\infty} (-)^r (m + r, r) q^{1/2r} Y_{m+r} \tag{17}$$

$$Y_n = \sum_{s=0}^{\infty} (n + s, s) q^{1/2s} X_{n+s}.$$

A particular set of functions satisfying the Hermitian relation

$$J_{m,n} = \int_{-\infty}^{\infty} \mathrm{Sech}^2 \left(\frac{1}{2}\pi x\right) F_m(ix) F_n(-ix) dx = 0 \quad m \neq n \tag{18}$$

is obtained by defining $F_n(u)$ to be a polynomial of the $n$th degree in $u$ such that

$$F_n\left(\frac{d}{dt}\right) \mathrm{Sech}\, t = \mathrm{Sech}\, t\, P_n(\tanh t) \tag{19}$$

where $P_n(t)$ is the Legrendre polynomial. It has already been shown[4] that

$$F_m(ix)\ \mathrm{Sech}\left(\frac{1}{2}\pi x\right) = \frac{1}{\pi}\int_{-\infty}^{\infty} e^{-ixz}\ \mathrm{Sech}\, z.\ P_m(\tanh z)\, dz \tag{20}$$

$$\therefore J_{m,n} = \frac{1}{\pi^2}\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} \mathrm{Sech}\, t.\ P_m(\tanh t)\, dt \int_{-\infty}^{\infty} e^{ix(t-z)} P_m(\tanh z)\, \mathrm{Sech}\, z\, dz$$

Changing the order of integration with respect to $x$ and $t$, which is permissible since

$$\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} e^{ixt} F_n(ix)\ \mathrm{Sech}\left(\frac{1}{2}\pi x\right) \mathrm{Sech}\, t\, P_n(\tanh t)\, dt$$

is absolutely convergent, we have

$$\begin{aligned} J_{m,n} &= \frac{2}{\pi}\int_{-\infty}^{\infty} \mathrm{Sech}^2 t.\ P_m(\tanh t)\, P_n(\tanh t)\, dt. \\ &= 0 \qquad m \neq n. \\ &= \frac{4}{\pi}\frac{1}{2n+1} \qquad m = n. \end{aligned} \tag{21}$$

[1] J. L. W. V. Jensen, *Nyt Tidsskr Math.*, **29**, 29 (1918). Jensen remarks that this generalized binomial coefficient plays an important part in Gauss' memoir "Summatie quarundam serierum singularium" (1811), in which the following relation is obtained

$$\begin{aligned} \sum_{r=0}^{n} (-)^r (n, r) &= (1 - q^n)(1 - q^{n-2}) \ldots (1 - q) \qquad n \text{ odd} \\ &= 0 \qquad n \text{ even} \end{aligned}$$

Jensen discusses the properties of the sums

$$\sum_{r=0}^{n} (-)^r q^{r^2} (n, r). \qquad \sum_{r=0}^{n} q^{r(s+1/2)} (n, r).$$

[2] R. Tambs Lyche, *Bull. Société mathématique France*, **55**, 102 (1927); *Compt. rend. t.* **186**, 1810 (1928); *Avhandlingar* Oslo (1928) No. 6; Forhandlingar Norske Videnskaber, Selskab 1, No. 35, 3 p. In the first paper Tambs Lyche shows that if

$$\phi_1(z) \equiv \phi(z) = qz + a_2z^2 + \ldots a_pz^p + \ldots$$

the $n$th iterated function is expressed formally by the series

$$\phi_n(z) = \phi[\phi_{n-1}(z)] = q^nz + a_2^{(n)}z^2 + \ldots a_p^{(n)}z^p + \ldots$$

where

$$a_p^{(n)} = \sum_{r=1}^{p-1} (-)^{p-r-1}\, q^{n-r+1/2(p-r)p-r-1}\, (n, r)(n-r-1, p-r-1)a_p^{(r)}$$

The notation $\begin{bmatrix} n \\ r \end{bmatrix}$ is preferable to either $(n, r)$ or $\begin{bmatrix} n \\ r \end{bmatrix}_q$ but we use $(n, r)$ here for convenience in printing. In 1897 F. H. Jackson used the notation $\begin{bmatrix} n \\ r \end{bmatrix}$ for the generalized binomial coefficient derived from $p$, where $p > 1$ but he also used it in a more general sense. See *Proc. London Math. Soc.* **28**, 475 (1897). The notation $(n)$ for $1 - q^n$ was suggested by A. Cayley and adopted by P. A. MacMahon. *Proc. London Math. Soc.*, ser. 2, **15**, 314 (1916). L. J. Rogers, *Ibid.*, **16**, 315 (1917), uses the notation $q_n$ for the same quantity and also writes $q_n!$ for the product $q_1\, q_2 \ldots q_n$.

[3] F. H. Jackson, *Proc. London Math. Soc.*, ser. 2, **1**, 63 (1904); **1**, 361 (1904); **2**, 192 (1904); **3**, 1 (1905); *Trans. Roy. Soc. Edinburgh*, **41**, 1, 105, 399 (1904–5); *Proc. Edinburgh Math. Soc.*, **22**, 80 (1904); *Proc. Roy. Soc. Lond.*, **76**, 127 (1905); *Messenger Math.*, ser. 2, **37**, 123 (1907); **38**, 57, 62 (1908); **39**, 26, 145 (1910); **40**, 92 (1910); **47**, 57 (1917); **50**, 101 (1920); **57**, 169 (1928); *Rendiconti Palermo*, **29**, 340 (1910); *Amer. Jour. Math.*, **32**, 305 (1910); *Proc. Roy. Soc. Edinburgh*, **30**, 378 (1910). Many of these papers deal with a generalization of the binomial theorem and generalizations of the functions of Legendre and Bessel all of which are more or less closely connected with the theory of the $q$-hypergeometric series. This series has been studied also by C. G. J. Jacobi, *Jour. Math.*, **32**, 197 (1846); E. Heine, *Ibid.*, **32**, 210 (1846); **34**, 285 (1847); **39**, 288 (1850); J. Thomae, *Ibid.*, **70**, 258 (1869); G. N. Watson, Cambridge, *Phil. Trans.*, **21**, 281 (1910). A new use of the generalized binomial coefficients in the theory of numbers has been found recently by I. Schur, *Sitzungsberichte preuss. Akad. Wiss.*, 145 (1933).

[4] H. Bateman, *Tôhoku Math. Jour.*, **37**, 23 (1933).

# THE EXPANSION OF THE UNIVERSE AND THE INTENSITY OF COSMIC RAYS

BY PAUL S. EPSTEIN

NORMAN BRIDGE LABORATORY OF PHYSICS, CALIFORNIA INSTITUTE OF TECHNOLOGY

Communicated December 4, 1933

1. It was pointed out by Zwicky[1] that the distance which a light quantum can travel is, perhaps, limited. Its frequency $\nu$ and energy $h\nu$ are affected by the astronomical red shift and may be reduced to the vanishing point if the distance is sufficiently large. Experimentally the dependence of the red shift upon the travel time $t - t_0$ follows very nearly (within the limits of error) a linear law up to $t - t_0 = 10^8$ years and amounts to $k = -\Delta\nu/\nu = 0.56 \times 10^{-9}$ per year. If it is permissible to extrapolate the linear law to larger distances, the light quantum will vanish for $t - t_0 = 1.8 \times 10^9$ years. All the radiation received by a terrestrial observer must, in this case, be derived from a space limited by this radius and this involves serious difficulties for the explanation of the rather high intensity of cosmic rays.

The question arises, therefore, to what extent the linear extrapolation of the red shift is justifiable and its analysis forms the main content of the present paper. The answer is that the embarrassingly short range of the cosmic rays is only a new disguise for the old difficulty of the short time scale of the expanding universe. In fact, the first rough estimates of the age of the universe were also made by a linear extrapolation of the red shift (which is of the nature of a Doppler effect) and, therefore, led to precisely the same figure $1.8 \times 10^9$ years. It is desirable, however, to bring out this connection with more detail. The shortness of the time scale was particularly emphasized by Lemaitre who showed that it can be avoided by assigning a proper value to Einstein's cosmological constant $\lambda$. Curiously, therefore, the high intensity of cosmic rays which we actually measure is, to a certain degree, an additional argument in favor of retaining the cosmological constant which some authors wished to discard. In this connection it will be necessary to discuss the different possible types of expansion of the universe and the time scales belonging to them. It is possible to do this in an extremely simple and general and, at the same time, more exhaustive way than it has been done heretofore.

2. We restrict ourselves, in this paper, to the case of the homogeneous (quasi-static) universe and take as our starting point Friedmann's form[2] of the non-static line element

$$ds^2 = -R^2(t)[d\chi^2 + \sin^2\chi(d\vartheta^2 + \sin^2\vartheta\, d\varphi^2)] + c^2dt^2, \qquad (1)$$

where $\chi$, $\vartheta$, $\varphi$ are the space coördinates which remain unchanged by the

expansion, $R(t)$ the variable radius of the world and $c$ a constant (velocity of light), and we wish to determine the energy of free motion in a geodesic. We can simplify the problem by choosing the coördinates $\vartheta$, $\varphi$ in a proper way. If the origin lies on the geodesic, we have $\varphi = \text{const.}$, $d\varphi = 0$, moreover we can define $\vartheta$ so as to make the geodesic lie in the equatorial sphere: $\vartheta = \pi/2$, $d\vartheta = 0$.

$$ds^2 = -R^2(t)\, d\chi^2 + c^2\, dt^2. \tag{1'}$$

In the case of light, $ds = 0$ and we obtain

$$d\chi = c\, R^{-1}\, dt, \tag{2}$$

or for the angular distance corresponding to a travel time from $t_0$ to $t$

$$\chi = c \int_{t_0}^{t} R^{-1}\, dt. \tag{3}$$

If at one end of this distance two signals are sent out at the interval $\Delta t_0$ from each other, they will arrive at the other end with the interval $\Delta t$. The ratio $\Delta t / \Delta t_0$ is obtained by differentiating with respect to $t_0$ the equation (3), where $\chi$ is independent of time,

$$\Delta t : \Delta t_0 = R : R_0, \tag{4}$$

if we write for short $R = R(t)$ and $R_0 = R(t_0)$, a formula which has been given in an equivalent form by Tolman[3] and others. In particular, we could consider the time intervals between two planes of equal phase of the light wave: In this case, the eq. (4) will give us the ratio between the periods which the light has in the points $\chi = \chi$ and $\chi = 0$. As the frequencies are reciprocal to the periods, the energies ($E = h\nu$) of a light quantum at the time of its arrival and at the time of its emission stand in the inverse ratio

$$E : E_0 = R_0 : R. \tag{5}$$

Almost equally simple are the conditions in the general case of any moving particle. The equations of its geodesic are obtained from (1') by Lagrangian differentiation

$$\frac{d}{ds}\left(R^2 \frac{d\chi}{ds}\right) = 0, \qquad c^2 \frac{d^2t}{ds^2} + R \frac{dR}{ds}\left(\frac{d\chi}{ds}\right)^2 = 0.$$

Eliminating $ds$ between them, we find an easily integrable equation which leads to

$$v = R \frac{d\chi}{dt} = c(1 + \alpha^2 R^2)^{-1/2}, \tag{6}$$

$\alpha$ being the constant of integration. If we take as the expression of the energy the usual $E = mc^2 dt/ds = mc^2(1 - v^2/c^2)^{-1/2}$, we obtain

$$\frac{E}{E_0} = \frac{R_0}{R}\left(\frac{1 + \alpha^2 R^2}{1 + \alpha^2 R_0^2}\right)^{\frac{1}{2}}. \qquad (7)$$

We see from (6) that for very fast particles of very high energy with velocities $v$ practically equal to the velocity of light $c$, the constant $\alpha$ must be so small as to make $\alpha^2 R^2$ negligible compared with 1. In this case (7) becomes identical with (5). If material particles form a part or the whole of the cosmic radiation, they unquestionably have velocities of this order. We can, therefore, apply the formula (5) to corpuscles in general without bothering about their physical nature (light quanta or matter). In the other limiting case, when $\alpha$ is very large, we find the approximation $E \propto (1 + 1/2\alpha^2 R^2)$. If we denote the constant of proportionality by $mc^2$,

$$E = mc^2(1 + 1/2\alpha^2 R^2), \qquad (5')$$

we recognize that the first term can be interpreted as the intrinsic, the second as the kinetic energy of the particle or body moving with a comparatively small velocity.

3. Although the law (5) of energy decrease of photons is not new, it seems that its main implications have never been discussed. It shows us that the energy of the corpuscles we observe at the present time is proportional to the radius of the universe at the time they were emitted. If the history of our universe is such that at some time in the past its radius was zero or nearly zero, then the case surmised by Zwicky is realized. The corpuscles which arrive with their energies completely exhausted are those which were sent out "in the beginning of the world" and their travel time $t - t_0$ is equal to the age of the universe. There are no signals possible from earlier states: The beginning of the world is at the same time the beginning of time.

Lemaitre expressed the hypothesis that the cosmic rays were produced just in these incipient stages of the universe when the density was high and matter was, perhaps, in a state quite different from what it is today. Radioactive materials with enormous molecules might have existed then whose breaking down produced correspondingly energetic $\alpha$-, $\beta$- and $\gamma$-rays. These rays may have been circulating round the universe ever since and they manifest themselves even today in the form of cosmic rays. Attractive as this picture is, the relations just stated present grave difficulties to it: The longer the hyper-radioactive rays circulate the more they are whittled down by red shift. In order to have the properties of cosmic rays, they must have started with energies millions of times higher than the very large ones we find in them now. If the hypothesis is given a less extreme form, to the effect that the super-radioactive processes went on not only in the initial but in the early stages of the universe,

it meets with less difficulty. Yet its advantages are somewhat decreased by the above considerations.

For short travel times $\Delta t$, we have only to substitute in (5) $E_0 = E - \Delta E$, $R_0 = R - \Delta R$, to obtain the well-known form of the red shift coefficient

$$k = -\frac{1}{E}\frac{\Delta E}{\Delta t} = \frac{1}{R}\frac{\Delta R}{\Delta t}. \tag{8}$$

We know the present value of this coefficient, and the question is how it was changing in former times. This will be answered if we come to know what function of time the radius $R$ of the universe is. The linear extrapolation of Zwicky's and others' amounts to assuming that $kR = \Delta R/\Delta t = \text{const}$. This means, if we plot $R$ against $t$ (Fig. 1), that the slope of the curve was at all times the same as it is at the present moment and the law is represented by a straight line, for instance, $AB$. The age of the universe would then be given by the segment $AC$. As this extrapolation is very dubious, we wish to investigate the problem somewhat closer. Much work has already been done on this subject: On the one hand, Friedmann[1] and Lemaitre discussed the possible types of expansion in the special case when the pressure can be neglected. On the other hand, Tolman[4] and Tolman and Ward[5] succeeded in excluding certain types of behavior as impossible. A general investigation, however, making no restrictions (but that the pressure remain positive) does not yet exist and shall be given in this and the next sections.[10]

The basis of this investigation is formed by the expressions of the energy-momentum tensor corresponding to Friedmann's line element (1) which in our notations have the form:

$$\left.\begin{aligned} -8\pi p &= (c^2 + \dot{R}^2 + 2R\ddot{R})R^{-2}c^{-2} - \lambda, \\ 8\pi\rho/3 &= (c^2 + \dot{R}^2)R^{-2}c^{-2} - \lambda/3, \end{aligned}\right\} \tag{9}$$

where $p$ is the pressure, $\rho$ the density and $\lambda$ the cosmological constant. The second of these eqs. gives us $\dot{R}$, while we obtain $\ddot{R}$ by taking the difference of the two

$$\dot{R} = c[(8\pi\rho + \lambda)R^2/3 - 1]^{1/2}, \tag{10}$$

$$\ddot{R} = c^2[\lambda - 4\pi(\rho + 3p)]R/3. \tag{11}$$

It will be useful to remark that, if we exclude negative pressures ($p \geqslant 0$), both $\rho$ and $p$ are necessarily decreasing functions of $R$. In fact, from (9) and (10) follows the identity

$$d(\rho R^3) + p\,dR^3 = 0, \tag{12}$$

(eq. of conservation) which is formally identical with the law of compression of an adiabatic fluid. When the pressure is so low as to be negligible ($p = 0$), this gives $\rho = C/R^3$; when $p$ is finite ($p > 0$), the rate

of decrease of $\rho$ is even faster. The pressure $p$ decreases, of course, with $\rho$ for reasons of stability.

4. The case of $\lambda \leqslant 0$ is now immediately disposed of: $R$ is, then, under all circumstances negative so that $\dot{R}$ decreases monotonically. Let us go backward in time from the point $B$ on curve I (Fig. 1) which represents the present state of the universe. The slope of the curve becomes steeper and steeper as is shown in the branch $BD$. Since $\rho$ is infinite for $R = 0$, the slope is vertical in the point $D$.[6] The age of the universe is represented by the segment $DC$ which is shorter than the segment $AC$ obtained by linear extrapolation. On the other hand, if we proceed from $A$ forward, the slope will decrease, becoming horizontal when $\dot{R}$ vanishes. This will occur when

$$Q(R) = (8\pi\rho + \lambda)R^2/3 - 1 = 0 \qquad (13)$$

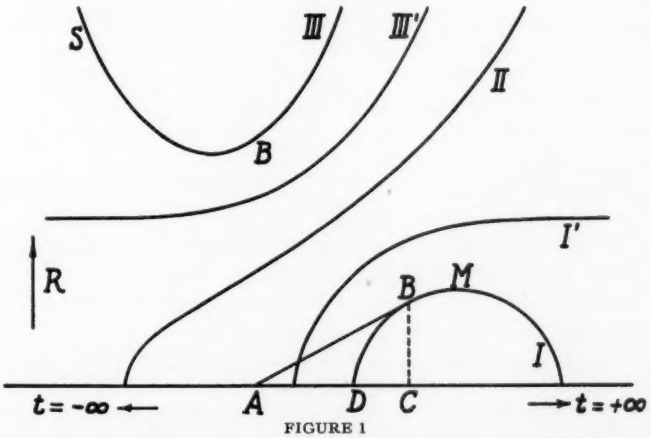


FIGURE 1

and, since the first term is a monotonic function of $R$, this eq. has always one and only one simple root. It corresponds to the maximum $M$ of the curve which is reached, according to well-known theorems,[7] in a finite time. Beyond it the slope becomes negative and the second half of the curve is symmetrical to the first.

In the case $\lambda > 0$, the character of the curve depends on the roots of both bracket expressions of the eqs. (10) and (11). Since $\rho + 3p$ is a monotonic function of $R$, the eq. $\lambda - 4\pi(\rho + 3p)$ has always one and only one root which we shall denote by $R_i$. For $R = R_i$ we have $\ddot{R} = 0$, so that it corresponds to a *point of inflexion* of our curve. On the other hand, in the expression $Q(R)$ of eq. (13) the term $8\pi\rho R^2/3$ is monotonically decreasing and $\lambda R^2$ is monotonically increasing. $Q(R)$ is infinite both for $R = 0$ and $R = \infty$ and has a minimum in between. We should expect

that the position of the minimum is $R = R_i$ (the root of the other bracket expression). This is, in fact, the case as is, easily, checked with the help of eq. (12). We have now to distinguish two cases:

(1) $Q(R_i) > 0$, the minimum value of $Q(R)$ is positive. Then $Q(R)$ and $\dot{R}$ are positive throughout and $R$ increases monotonically as shown in curve II of figure 1. The slope is vertical for $R = 0$ and decreases until the point of inflexion $R = R_i$ is reached. From then on, it increases and becomes again vertical for $R = \infty$.

(2) $Q(R_i) < 0$. In this case, $Q(R)$ has two roots $R_1$ and $R_2$ so that $R_1 < R_i$ and $R_2 > R_i$. The radius $R$ cannot assume values between $R_1$ and $R_2$ because this would make the rate of expansion $\dot{R}$ imaginary. We have, therefore, two alternatives: (Case 2*a*, when $R < R_1$ and, *a fortiori*, $R < R_i$). The acceleration $\ddot{R}$ is, then, permanently negative so that the conditions are quite the same as in the case $\lambda \leqslant 0$ and we are led again to a curve of the type I. Its maximum $M$ corresponds to $R = R_1$. (Case 2*b*, when $R > R_2$ and, *a fortiori*, $R > R_i$). The second derivative $R$ is then permanently positive and the curvature opposite to that of curve I. We find then a curve of the type III which goes into infinity for $t = \pm\infty$ and touches a minimum for $R = R_2$. Its slope tends toward the vertical at both extremities.

In the intermediate case $Q(R_i) = 0$, $R_i$ is a double root of $Q(R)$, so that the time of approach of this value becomes logarithmically infinite. The horizontal $R = R_i$ is, then, an asymptote which can be approached either from below, as in curve I′, or from above, as in curve III′.

5. By the types of figure 1 all the possibilities are exhausted. We have said that this applied to the general case, in the sense that it will be true for any law of change of pressure applicable to fluids (i.e., $p$ must be positive and increase with $\rho$). If the law is such that $\rho$ becomes infinite already for a finite volume, smaller volumes have no physical reality and the axis of abscissae of figure 1 corresponds not to $R = 0$ but to a finite minimum value of the radius without any change in the type of the curves.

It must not be forgotten, however, that the Einstein theory of continuous matter which leads to Friedmann's eq. (1) is based on the consideration of a system of material points. It is, so to speak, the theory of a gravitating perfect fluid, in which the size of the particles and non-gravitational forces between them are omitted. It would be futile to try to improve the accuracy by discussing laws of pressure which take the interaction into consideration because they are outside the validity of the fundamental equation. But this does not impair the theory as a good approximation to the real conditions. Because of the present very low density of matter even the big agglomerations are sufficiently wide apart. On the other hand, if the volume of the universe should become (or have been) considerably smaller, these large units will dissolve and so the

process will go on by successive stages. We can, therefore, expect important deviations from the law of perfect fluids only for extremely small radii $R$.

The equation of state for our perfect fluid follows directly from our energy expressions (5) and (5′). In the stellar space we have, on one hand, bodies and particles moving with astronomical velocities which are small compared with the velocity $c$ of light; on the other hand photons and other corpuscles having quite or almost the velocity $c$. For the first group applies the energy expression (5′), for the second (5). If we take the sum of the energies of all these bodies, the total energy of the universe $U$ will consist of three terms: the intrinsic energy of the slow moving elements $U_i$ = const., their kinetic energy $U_k \propto 1/R^2$, the radiant energy (including the total energy of the $\alpha$- and $\beta$-rays) $U_r \propto 1/R$. Differentiating the logarithms of $U_k$, $U_r$, we find the expressions $dU_k/dR = -2U_k/R$, $dU_r/dR = -U_r/R$. If $V$ is the volume of the universe, $U = \rho V \propto \rho R^3$. We can, therefore, write

$$\frac{d(\rho_k R^3)}{dR} = -2\rho_k R^2, \qquad \frac{d(\rho_r R^3)}{dR} = -\rho_r R^2. \tag{14}$$

Because of the linear character of eq. (12), the two types of particles produce their separate partial pressures: $p = p_k + p_r$, and this eq. separates into two with the indices $k$ and $r$. Comparing them with (14), we obtain

$$p_k = \frac{2}{3}\rho_k, \qquad p_r = \frac{1}{3}\rho_r \tag{15}$$

precisely the same relations as in the non-relativistic theory.

It follows from these considerations that when the universe expands into infinity ($R \to \infty$), both $U_k$ and $U_r$ vanish and only $U_i$ can remain finite. Conversely, when it contracts to nothing ($R \to 0$), not only the energies $U_k$, $U_r$ but even the densities $\rho_k$, $\rho_r$ become infinite.

6. We return now to the question about the maximum travel time of photons and other fast corpuscles. As was shown in section 3, it coincides with the age of the universe represented in figure 1 by the projection of the expansion curve upon the time axis from the intercept to a point B giving the present state. It must be compared to the subtangent at the point B which is known from the red shift measurements to be equal to $1.8 \times 10^9$ years. There follows from the analysis of section 4 that, in the case of negative or vanishing cosmological constant, the range of all corpuscles is necessarily smaller than $1.8 \times 10^9$ years. If $\lambda$ is positive, the range depends on the value of this constant and of $R(t)$ and can assume all possible values up to infinity.

Lemaitre has already pointed out that the types of motion giving less than $1.8 \times 10^9$ years must be excluded because the geological time scale of the earth is of this order of magnitude. The difficulty of explaining the large intensity of cosmic rays, if their range is limited, point in the same direction. To what extent do astronomical data permit us to decide the question what type of expansion our universe is actually undergoing? The observational evidence is very scant: On the one hand, we have the Hubble-Humason red shift measurements which give $k/c = 5.6 \times 10^{-10}$ (years)$^{-1} = 5.8 \times 10^{-28}$ cm.$^{-1}$ On the other, Hubble's determination of the minimum density $8\pi\rho = 8\pi 10^{-30}$ g. cm.$^{-3} = 1.86 \times 10^{-57}$ cm.$^{-2}$ This value is derived for the luminous part of matter and does not include the dark part. The actual density may, therefore, be considerably above this minimum. Eliminating $\dot{R}$ from the eqs. (8) and (10) we find

$$(\lambda + 8\pi\rho - 3k^2c^{-2})R^2 = 3, \tag{16}$$

so that the parenthesis must be positive and we have the inequality

$$\lambda \geqslant 3k^2c^{-2} - 8\pi\rho. \tag{17}$$

Now $3k^2c^{-2} = 1.0 \times 10^{-54}$ and Hubble's minimum for $8\pi\rho$ is more than five hundred times smaller. If it is possible to assume that the true density is not higher than about 100 times this minimum, $\lambda$ will be determined, essentially, by the first term on the right side and will be at least $10^{-54}$. This value of $\lambda$ coincides with that arrived at by Lemaitre on the basis of somewhat different considerations. Under the conditions assumed above we find $\lambda - 4\pi(\rho + 3p) > 0$ (since the pressure $p$ is negligible compared with $\rho$) and this means, according to section 4, that in the present state of the universe the curvature of the expansion curve is convex toward the time axis. This leaves us the choice between the types II and III. With the help of our formulas and neglecting $p$ we obtain the criterion $R \gtrless (4\pi\rho)^{-1/3}\lambda^{-1/6}$ or $R \gtrless 1.0 \times 10^{28}$ cm. (to $2 \times 10^{27}$ cm.) the upper sign referring to the type II and the lower to III. Both possibilities are compatible with all known facts and a decision between the two types cannot be made. Even worse than that, it is possible that the density $\rho$ is thousands of times higher than the lower limit given by Hubble, in which case even the type I would be permissible if there were not the difficulty of the short time scale.

Yet it must be said that the main argument which leads to a very high density $\rho$ is to be taken with some caution. It is based on the fact that the irregular velocities of stars and nebulae are unusually high. Their explanation as the result of a Newtonian attraction by nebulae would require very large mutual forces and correspondingly large masses. However, the analogy between a classical system and the expanding universe

is very imperfect. The observable kinetic energy in the Einstein-Friedmann universe is not the analogue to the kinetic energy of a point system subject to Newtonian forces because there is still the expansion of the radius. As we have seen in sections 2 and 5, it increases according to an entirely different law. We wish to illustrate this for the case of an expansion of the type given by curve III (Fig. 1) as this type has some remarkable properties which are well worth discussing. It was pointed out that in the initial state of this kind of a universe ($R = \infty$) both the total and the individual kinetic and radiant energies are infinitely small. A photon of the energy $E_S$ could have been emitted only when the radius was already finite as represented, say, by the point $S$ on curve III. As the universe continues to shrink the energy of the photon increases, according to our formula (5), in reciprocal proportion to the radius. After the minimum is passed, it decreases again and assumes the value $E_B$ in the point $B$ representing the present state of the universe. If $E_S$ is the energy of a radioactive $\gamma$-ray, $E_B = (R_S/R_B)E_S$ may be well that of a cosmic ray, provided the emission took place a very long time back. In a similar way the kinetic energy of a slowly moving celestial body will first increase and then decrease according to formula (5′), i.e., reciprocally to the square of the radius, so that the velocity will again follow the law $v_B = (R_S/R_B)v_S$. Comparatively small velocities $v_S$ due to accidental local causes will, therefore, be greatly increased.

The type of expansion III seems little attractive and we do not strongly advocate it, but it is an amusing fact that it could explain two great puzzles: The enormous energy of the cosmic rays and the high irregular velocities of stars. It is true that we did not take into account the collisions between celestial bodies: however, such collisions would only partially reduce the kinetic energy. In the case of the cosmic rays, there arises the question why we do not observe visible light coming from equally large distances. This might be explained by a difference in absorption: We shall see in the next section that it is possible to assume that light is completely absorbed in a comparatively short travel path, while cosmic rays could have a much longer range. On the other hand, apart from other considerations, it is against the type III that the necessary assumption of vanishing kinetic and radiant energies in the initial state seems rather artificial.

Independently of any assumption about the type of expansion holds the following statement: As long as collisions and interactions are neglected the mean kinetic energy of celestial bodies is determined by a separate constant of integration and cannot be inferred from the density and radius of the universe. This fact is more striking in the relativistic theory than in the Newtonian: If in the former the kinetic energy is strictly zero at one time, it vanishes at all times.

7. In this section we shall discuss a few questions specifically relating to the motion of photons. We have mentioned the hypothesis that the cosmic rays were emitted in the early stages of the evolution of our universe. How many times did they circle around it during their long history? The angular distance $\chi$ which they cover follows from eqs. (2), (10) and (13) as

$$d\chi = R^{-1}Q^{-\frac{1}{2}}\,dR, \qquad \chi = \int_{R_0}^{R} R^{-1}Q^{-\frac{1}{2}}\,dR. \tag{18}$$

In the particular case considered by Einstein ($\lambda = 0$ and $p = 0$, $\rho = C/R^3$) this gives, for $R_0 = 0$, the eq. of a cylcoid $R = \frac{1}{2}R_1(1 - \cos\chi)$, $ct = \frac{1}{2}R_1(\chi - \sin\chi)$. This means that a pulse of light emitted in the "beginning of the world" ($R = 0$) comes back exactly at the "end of the world" ($R = 0$, again). If we take into consideration the finite values of the pressure $p$ but leave $\lambda$ unchanged, it is easy to see that the angular distance covered by the rays will be even smaller. This fact is very remarkable as it shows that the different parts of the universe act as one whole, with respect to the expansion, long before they could have received light signals from one another. Its closer investigation is contemplated by the writer since it should be very instructive. On the other hand, a positive cosmological constant increases the angular range of light but, for the ascending part of curve I, not very considerably. Even in the limiting case $I'$, for a state in which the radius $R$ has reached one half of its final value, the maximum angular range turns out to be only about 60°. We find similar conditions in the case of the curve III: If $\rho$ can be neglected beside $\lambda$, the angular distance which a ray of light covers during the whole evolution (from $t = -\infty$ to $t = +\infty$) is $\pi$ or one-half circle. As $\rho$ becomes more important, this distance increases and becomes infinite in the limiting case III′. In the case of the type II, it is also short except when the curve approximates the limiting case III′. This shows that the life of the universe must be considered, in general, as extremely short. Only for laws of expansion in the neighborhood of the curve III′ could we expect some measure of equilibrium. Vice versa, the fact that we actually do not find equilibrium between matter and radiation, etc.,[8] is not at all surprising.

Let us now compute the total intensity of radiation received on the surface of the earth. Suppose that a star at the distance $\chi$ or ($t - t_0$ in travel time) from us emits in the time $dt_0$ the energy $du_0 = \epsilon_0 dt_0$. By the time it reaches us it is spread out over a sphere of the surface $4\pi R^2 \sin^2\chi$ and, moreover, attenuated by the red shift according to our formula (5).

The energy received per unit area is, therefore, $du = \epsilon_s R_0 dt_0/4\pi R^3 \sin^2 \chi$. The intensity received from that star becomes $I_s = du/dt = \epsilon_s R_0^2/4\pi R^4 \sin^2 \chi$. If we wish to consider not a single star but all the radiating matter in a spherical layer between $\chi$ and $\chi + d\chi$, we must introduce a coefficient of emission $\epsilon(\chi)$ referred to unit mass. The total mass of the layer in question is $4\pi\rho_0 R_0^3 \sin^2 \chi d\chi$ and, to obtain the intensity of radiation due to it, we have to replace $\epsilon_s$ by the mass multiplied by $\epsilon(\chi)$. Integrating with respect to $d\chi$, we find the total intensity (with neglect of absorption)

$$I = \frac{1}{R^4} \int_0^\chi \epsilon(\chi) \rho_0 R_0^5 \, d\chi, \tag{19}$$

where the $\chi$ in the limit of the integral is the angular distance which a ray travels from the beinning of the universe to the present moment. If it is possible to neglect the pressure, the total mass remains constant: $\rho_0 R_0^3 = \rho R^3$, and the expression reduces to

$$\begin{aligned} I &= \frac{\rho}{R} \int_0^\chi \epsilon(\chi) R_0^2 d\chi = \frac{\rho}{R} \int_{\chi=0}^{t} \epsilon(\chi) R_0 dt_0 \\ &= \frac{\rho}{R} \int_{\chi=0}^{R} \epsilon(\chi) R_0 \rho_0^{-1/3} dR_0. \end{aligned} \tag{20}$$

If we make $\epsilon(\chi)$ constant, the integral becomes divergent for the type III: We have already pointed out (section 5) that, in this case, no radiation can take place while the radius is infinite. As to the types I and II, the intensity is smaller than the expression which results from $R_0$ being replaced by $R$, i.e., $I < \rho\epsilon t$. And, in particular, for the type I, it is *a fortiori* smaller than the values $\rho\epsilon \times 1.8 \times 10^9$ (years) obtained by Zwicky from linear extrapolation.

Zwicky's argument by which he shows that, provided the emission of cosmic rays is a general property of matter,[9] the range of $2 \times 10^9$ light years is much too short to account for their intensity remains, therefore, entirely valid. The explanation may be either that the range is longer (that is to say, the expansion of the universe is not of the type I), or the cosmic rays are emitted under special conditions. One possibility is that suggested by Lemaitre that they are the messengers of some super-radioactive materials which existed once but have long since disappeared in the solar system. Mathematically this would mean that $\epsilon(\chi)$ in eq. (20) has to increase with $\chi$ much more rapidly than $R_0$ decreases. We remind, however, of the remarks made on this question in section 3. It was argued that if we admit for the cosmic rays very long ranges, say, of the order of $10^{11}$ light years, we must admit them also for visible light and

this would be in contradiction with the comparative darkness of the night sky. This objection would not hold, however, if the main part of the absorbing matter in the interstellar spaces were in the form of dust. As an example, let us consider the case of $N$ dust grains per cm.$^3$ each of the radius $a = 10^{-2}$ cm., consisting of some dark material which absorbs most of the light falling on it and has the specific gravity $d$. A pencil of light rays would be completely absorbed after traveling the distance $L_v = 1/\pi Na^2$ cm. On the other hand, the total mass of the dust particles is $4\pi a^3 Nd/3$ per cm.$^3$, and if a cosmic ray has the range $l = 100$ cm. in water, it will have in the dust range $L_c = 3l/4\pi a^3 Nd$. The ratio $L_c/L_v = 3l/4ad$ is of the order 1000 if $d$ is about 7.5. Visible light can, therefore, easily have an absorption thousands of times higher than cosmic rays so that the bulk of the two kinds of radiation falling upon the earth is drawn from quite different parts of the universe. If we take $L_v = 10^{10}$ light years, the density of disperse dark matter in this example works out to be $10^{-29}$ g./cm.$^3$, an entirely reasonable value.

[1] F. Zwicky, *Phys. Rev.*, **43**, 147 (1933).

[2] A. Friedmann, *Zeitschr. Physik*, **10**, 377 (1922).

[3] R. C. Tolman, *Proc. Nat. Acad. Sci.*, **16**, 320 (1930).

[4] R. C. Tolman, *Phys. Rev.*, **38**, 1758 (1931).

[5] R. C. Tolman and M. Ward, *Ibid.*, **39**, 835 (1933).

[6] Compare the remarks on this point in the beginning of section 5.

[7] Compare L. Charlier, "Die Mechanik des Himmels," Vol. I, pp. 85–90; Leipzig (1902).

[8] Compare R. C. Tolman, *Proc. Nat. Acad. Sci.*, **17**, 153 (1931).

[9] As an average over very large material systems, not for every kind of matter separately.

[10] The writer discovered a paper by H. P. Robertson (*Rev. of Modern Physics*, **5**, 62, 1933) which contains a classification of types of expansion similar to that of our section 4.

# SCIENTIFIC WORK IN THE "CENTURY OF PROGRESS" STRATOSPHERE BALLOON

BY ARTHUR H. COMPTON

RYERSON PHYSICAL LABORATORY, UNIVERSITY OF CHICAGO

Communicated December 7, 1933

On November 20, 1933, the balloon *Century of Progress* rose from Akron to an official altitude of 18,665 meters (barometer 49.5 mm.), and descended eight hours later near Bridgeton, New Jersey. During the flight experiments were carried on to determine: 1. The ionization by cosmic rays using an unshielded ionization chamber. 2. The ionization by cosmic rays in a chamber shielded by 7.5 cm. of lead shot, or the equivalent of 5 cm. of solid lead. 3. The number and directional distribution of cosmic rays with a double Geiger-Müller coincidence counter. 4. The photographic action of cosmic rays on x-ray films. 5. The transmission of the atmosphere for the solar spectrum, using a quartz spectrograph. 6. The state of polarization of the scattered light from the sky, using a crystal of calcite painted with uniformly spaced black bands, so as to compare visually the intensity of the ordinary and extraordinary rays. 7. The color of the sky, by comparison with color charts. 8. The composition of the air at high altitudes, by taking samples for subsequent laboratory analysis. 9. The viability of certain spores exposed to the low pressure and temperature of the stratosphere. 10. The intensity, range, and freedom from static of radio signals originating at high altitudes. 11. The photographic "visibility" of the earth, using ordinary and infra-red light. At the last moment it was not possible to include an experiment planned by A. D. Boyer of the University of Chicago for testing the action of cosmic rays on fruit flies, nor one by D. S. Hsiung of the University of Chicago for repeating the Bothe-Kohlhörster experiment with cosmic rays at high altitudes.

Both sets of apparatus for measuring the ionization by cosmic rays were self-recording, and gave satisfactory records. Due in part to an unexpected rapid rotation of the balloon, the program of directional experiments could not be completed, though valuable data on the number of rays were obtained. The rotation of the gondola also made it impossible to use the adjustable support for the quartz spectrograph, and attempts to secure spectrograms by holding the instrument in the hands gave unsatisfactory results.

The skylight at 90 degrees from the sun was completely polarized within an estimated limit of about 5 per cent. Close to the horizon the sky was white, due to haze and clouds. This shaded into a steel blue, and then into a saffron yellow, due apparently to light scattered from distant cirrus

clouds, and traversing a long path through the atmosphere. Above this highest layer of cloudy or hazy air, the color shaded rapidly through green to a deep blue. This blue showed no tinge of purple or violet hue. These observations were confirmed by the use of a hand spectroscope.

The radio signals were transmitted with surprising clarity on a wavelength of 19.7 meters. Immediately upon leaving the earth the oscillation frequency of the transmitter changed by about 30 per cent, due doubtless to change in capacity. Whereas close to the ground almost 100 per cent modulation was necessary to transmit satisfactory signals, at high altitudes 30 per cent modulation was sufficient. The signals were received directly at Chicago, New York and intermediate stations.

The rate of drift of the balloon while in the atmosphere was about 45 knots. This is considerably greater than occurred in previous stratosphere ascensions, though because of the different time of year is not surprising.

The temperature of the top of the balloon in the stratosphere rose to $-2°$C., though the air temperature was about $-55°$C. As the sun came close to the horizon the bag cooled rapidly, causing a descent which attained a speed of 18 meters per second. Of physiological interest is the fact that when the gondola was opened at an elevation of 8.1 kilometers (less than $^1/_3$ of an atmosphere pressure) none of the symptoms of dizziness, nausea, fatigue, etc., familiar at high altitudes were noticed, though both observers were engaged in heavy physical labor. Two explanations are apparent, first, the short period of exposure to the low pressure because of the rapid descent of the balloon, and, second, a possible high concentration of oxygen in the gondola before the ports were opened.

More detailed accounts of the results of the various experiments will be published by the individuals most directly responsible for them.

The balloon and gondola were built after the general design used by one of us in previous stratosphere flights, but with certain modifications demanded by the experiments and conveniences of navigation. The capacity of the balloon was somewhat greater, about 17,000 cubic meters, and the gondola, built of 95 per cent magnesium (Dowmetal), was somewhat lighter, 160 kilograms. The balloon was filled with 99.9 per cent hydrogen. The balloon, gondola, two men and essential equipment weighed a total of 1700 kilograms. In addition, 1800 kilograms of ballast, in the form of sand and lead dust, were carried at the start of the flight.

The barometer equipment consisted of two sealed aneroid barographs mounted outside the gondola, a double U-tube mercury in glass barometer inside the gondola, connected by rubber and copper tubing to the outer air, and two aneroid altimeters to determine the pressure within the gondola.

We wish to thank the Century of Progress Exposition for organizing the flight, the Chicago Daily News, the National Broadcasting Company, the Dow Chemical Company and the Union Carbon and Carbide Company for bearing the chief part of the cost of the flight. Of the many who have given valuable help in making the enterprise a success, we wish especially to thank the Goodyear-Zeppelin Corporation for their care in building the balloon, and the use of their personnel and their airship dock for handling the balloon previous to the flight; Admiral E. J. King, Chief of the Bureau of Aeronautics, U. S. N., for supplying essential navigating equipment; the Wm. Gaertner Scientific Corporation for supplying the quartz spectrograph; the U. S. Bureau of Standards for their assistance in connection with the barographs and other control instruments both before and after the flight; the U. S. Weather Bureau for their wholehearted assistance; the Navy Bureau of Construction and Repair for supplying submarine breathing equipment; the Hydrographic Office for navigational charts; the Bausch and Lomb Optical Company for loan of optical instruments, The Eastman Kodak Company for supplying some special photographic equipment, and Mr. T. J. O'Donnell, foreman of the Ryerson Laboratory shop, for his help in the final adjustments and tests of the gondola.

Luis Alvarez, University of Chicago
A. H. Compton, Univ. of Chicago
C. L. Fordney, U. S. Marine Corps
H. G. Gale, Univ. of Chicago
G. F. Smith, Univ. of Illinois
F. C. Meier, Dept. of Plant Pathology, U. S. Dept. of Agriculture
R. A. Millikan, Calif. Inst. of Technology
G. S. Monk, Univ. of Chicago
A. Piccard, Univ. of Brussels
J. Piccard, Marshallton, Del.
T. G. W. Settle, U. S. Navy
R. J. Stephenson, Univ. of Chicago

# ANALYSIS OF NON-COMMUTING VECTORS WITH APPLICATION TO QUANTUM MECHANICS AND VECTOR CALCULUS

BY G. H. SHORTLEY AND G. E. KIMBALL*

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Communicated November 29, 1933

In quantum mechanics it is necessary to consider in detail the properties of certain vectors whose components are linear operators. Two components of the same or different vectors do not in general commute. The manipulation of such vectors is facilitated by a set of formulas modified from those of the usual vector analysis.

If $\boldsymbol{A}$ and $\boldsymbol{B}$ are two non-commutative vectors ($\boldsymbol{A} = A_x i + A_y j + A_z k$), the commutators $[A_m, B_n] = A_m B_n - B_n A_m$ connecting the components of $\boldsymbol{A}$ and $\boldsymbol{B}$ transform like the components of a dyadic; hence it is convenient to write

$$[\boldsymbol{A}, \boldsymbol{B}] = [A_x, B_x]\, ii + [A_x, B_y]\, ij + \ldots . \tag{1}$$

This dyadic is not the dyadic $\boldsymbol{AB} - \boldsymbol{BA}$, but is related to it as follows:

$$[\boldsymbol{A}, \boldsymbol{B}] = \boldsymbol{AB} - (\boldsymbol{BA})_C = \boldsymbol{AB} - \boldsymbol{BA} - (\boldsymbol{B} \times \boldsymbol{A}) \times \mathfrak{I}, \tag{2}$$

where $\mathfrak{I}$ is the unit dyadic $ii + jj + kk$, and $_C$ indicates the conjugate (transposed) dyadic.[1] Hence in general $[\boldsymbol{B}, \boldsymbol{A}] \neq -[\boldsymbol{A}, \boldsymbol{B}]$, but from (2) we see that

$$[\boldsymbol{B}, \boldsymbol{A}] = -[\boldsymbol{A}, \boldsymbol{B}]_C. \tag{3}$$

In a similar way, the commutators of a scalar operator $P$ with the components of a vector $\boldsymbol{A}$ form a vector:

$$[P, \boldsymbol{A}] = -[\boldsymbol{A}, P] = [P, A_x]\, i + [P, A_y]\, j + [P, A_z]\, k. \tag{4}$$

With these definitions we may derive a set of rules for the manipulation of such vectors. All associative and distributive properties of ordinary vector analysis are preserved; other elementary operations must be performed in accordance with the formulas

$$\begin{aligned}
\boldsymbol{A}\cdot\boldsymbol{B} \times \boldsymbol{C} &= \boldsymbol{A} \times \boldsymbol{B}\cdot\boldsymbol{C}, && (5)\\
\boldsymbol{A} \times (\boldsymbol{B} \times \boldsymbol{C}) &= \boldsymbol{BA}\cdot\boldsymbol{C} - \boldsymbol{A}\cdot\boldsymbol{BC} - [\boldsymbol{B}, \boldsymbol{A}]\cdot\boldsymbol{C} && (6)\\
&= \boldsymbol{A}\cdot\boldsymbol{CB} - \boldsymbol{A}\cdot\boldsymbol{BC} - \boldsymbol{A}\cdot[\boldsymbol{C}, \boldsymbol{B}], && (7)\\
(\boldsymbol{A} \times \boldsymbol{B}) \times \boldsymbol{C} &= \boldsymbol{BA}\cdot\boldsymbol{C} - \boldsymbol{AB}\cdot\boldsymbol{C} - [\boldsymbol{B}, \boldsymbol{A}]\cdot\boldsymbol{C} && (8)\\
&= \boldsymbol{A}\cdot\boldsymbol{CB} - \boldsymbol{AB}\cdot\boldsymbol{C} - \boldsymbol{A}\cdot[\boldsymbol{C}, \boldsymbol{B}], && (9)\\
\boldsymbol{A}\cdot\boldsymbol{B} &= \boldsymbol{B}\cdot\boldsymbol{A} + [\boldsymbol{A}, \boldsymbol{B}]_s, && (10)\\
\boldsymbol{A} \times \boldsymbol{B} &= -\boldsymbol{B} \times \boldsymbol{A} + [\boldsymbol{A}, \boldsymbol{B}]_x, && (11)
\end{aligned}$$

where $_s$ and $_x$ indicate the scalar and vector obtained by inserting dots and crosses, respectively, between the two vectors of each dyad in a dyadic—the scalar and vector of the dyadic as defined by Gibbs.

The following commutation rules are very useful:

$$[\boldsymbol{A}\cdot\boldsymbol{B}, \boldsymbol{C}] = \boldsymbol{A}\cdot[\boldsymbol{B}, \boldsymbol{C}] - [\boldsymbol{C}, \boldsymbol{A}]\cdot\boldsymbol{B}, \tag{12}$$

$$[\boldsymbol{A}\times\boldsymbol{B}, \boldsymbol{C}] = \boldsymbol{A}\times[\boldsymbol{B}, \boldsymbol{C}] - ([\boldsymbol{C}, \boldsymbol{A}]\times\boldsymbol{B})_C, \tag{13}$$

$$[\boldsymbol{A}\times\boldsymbol{B}, P] = \boldsymbol{A}\times[\boldsymbol{B}, P] + [\boldsymbol{A}, P]\times\boldsymbol{B}. \tag{14}$$

As an example of the application of these rules to quantum mechanics we may derive the commutation rules connecting the position $r$ and momentum $p$ of an electron with the angular momentum $L = r \times p$, and in particular show that $r^2$ and $p^2$ commute with $L$. The basic commutation rules of Heisenberg are

$$[r, r] = 0, \quad [p, p] = 0, \quad [r, p] = -[p, r] = ih\mathfrak{I}. \tag{15}$$

From (13) we find that

$$[L, r] = [r \times p, r] = r \times [p, r] = -ihr \times \mathfrak{I} = [r, L], \tag{16}$$

$$[L, p] = [r \times p, p] = (ih\mathfrak{I} \times p)_C = -ihp \times \mathfrak{I} = [p, L], \tag{17}$$

since $\boldsymbol{A} \times \mathfrak{I} = \mathfrak{I} \times \boldsymbol{A} = -(\mathfrak{I} \times \boldsymbol{A})_C$ for any vector $\boldsymbol{A}$. From (12) we have

$$[r^2, L] = r\cdot[r, L] - [L, r]\cdot r = -ih(r\cdot\mathfrak{I} \times r - r \times \mathfrak{I}\cdot r) = 0$$

$$[p^2, L] = 0.$$

Aside from their applicability to the quantum mechanical operators, these formulas are of interest in ordinary vector analysis in that they furnish a rigorous method of handling the operator $\nabla = \boldsymbol{i}\,\partial/\partial x + \boldsymbol{j}\,\partial/\partial y + \boldsymbol{k}\,\partial/\partial z$ without the necessity of the usual calculation by components. The properties of $\nabla$ are completely given by the following set of commutators, where $u$ and $v$ represent ordinary *commutative* vectors, $u$ and $v$ ordinary scalars:

$$[\nabla, \mathbf{u}] = \{\nabla\mathbf{u}\}, \tag{18}$$

$$[\mathbf{u}, \nabla] = -\{\nabla\mathbf{u}\}_C = -\{\nabla\mathbf{u}\} - \{\nabla \times \mathbf{u}\} \times \mathfrak{I}, \tag{19}$$

the last equation following from (3) and (2);

$$[\nabla, u] = -[u, \nabla] = \{\nabla u\}. \tag{20}$$

By convention, we let $\nabla$ operate on everything which follows it unless $\nabla$ and the operand are set off by braces as on the right of the above relations. These commutators turn out then to be not operators, but ordinary dyadics and vectors.

In terms of these commutators and the rules (5) to (11) we may, for example, derive with ease all the formulas on page 157 of Gibbs-Wilson.[2]

without referring to the original definition of $\nabla$. The first two formulas depend only on the distributive property. The third follows from (20):

$$\nabla uv = u\nabla v + \{\nabla u\}v ;$$

the next two from (10), (11) and (18)

$$\nabla \cdot \mathbf{v}u = \mathbf{v}\cdot\nabla u + \{\nabla\cdot\mathbf{v}\}u$$
$$\nabla \times \mathbf{v}u = -\mathbf{v}\times\nabla u + \{\nabla\times\mathbf{v}\}u.$$

The gradient of $\mathbf{u}\cdot\mathbf{v}$ is given by the calculation:

$$\begin{aligned}\nabla\mathbf{u}\cdot\mathbf{v} &= \mathbf{u}\nabla\cdot\mathbf{v} + [\nabla,\mathbf{u}]\cdot\mathbf{v} + (\mathbf{u}\times\nabla)\times\mathbf{v} && \text{—from (2)}\\ &= \mathbf{u}\nabla\cdot\mathbf{v} + \{\nabla\mathbf{u}\}\cdot\mathbf{v} + \mathbf{u}\cdot\mathbf{v}\nabla - \mathbf{u}\nabla\cdot\mathbf{v} - \mathbf{u}\cdot[\mathbf{v},\nabla] && \text{—from (18) and (9)}\end{aligned}$$

(an expression such as $\mathbf{u}\cdot\mathbf{v}\nabla$ vanishes—$\nabla$ is to be considered as operating on unity; this vanishing follows formally from (19))

$$= \mathbf{v}\cdot\{\nabla\mathbf{u}\}_C - \mathbf{u}\cdot[\mathbf{v},\nabla]$$

(from ordinary vector analysis $\Phi\cdot\mathbf{v} = \mathbf{v}\cdot\Phi_C$)

$$= \mathbf{v}\cdot\{\nabla\mathbf{u}\} + \mathbf{v}\times\{\nabla\times\mathbf{u}\} + \mathbf{u}\cdot\{\nabla\mathbf{v}\} + \mathbf{u}\times\{\nabla\times\mathbf{v}\} \quad \text{—from (19)}$$

Finally the divergence and curl of $u \times v$ are obtained as follows:

$$\begin{aligned}\nabla\cdot\mathbf{u}\times\mathbf{v} &= \nabla\times\mathbf{u}\cdot\mathbf{v} && \text{—from (5)}\\ &= -\mathbf{u}\times\nabla\cdot\mathbf{v} + \{\nabla\times\mathbf{u}\}\cdot\mathbf{v} && \text{—from (11)}\\ &= -\mathbf{u}\cdot\nabla\times\mathbf{v} + \mathbf{v}\cdot\nabla\times\mathbf{u}; && \text{—from (5)}\\ \nabla\times(\mathbf{u}\times\mathbf{v}) &= \nabla\cdot\mathbf{v}\mathbf{u} - \nabla\cdot\mathbf{u}\mathbf{v} - \nabla\cdot[\mathbf{v},\mathbf{u}] && \text{—from (7)}\\ &= \mathbf{v}\cdot\nabla\mathbf{u} + \{\nabla\cdot\mathbf{v}\}\mathbf{u} - \mathbf{u}\cdot\nabla\mathbf{v} - \{\nabla\cdot\mathbf{u}\}\mathbf{v}. && \text{—from (10)}\end{aligned}$$

[1] See Gibbs-Wilson, *Vector Analysis*, Chapter V, for a complete discussion of the ordinary theory of dyadics. It should be pointed out that in combining non-commutative vectors, the order of the factors must be preserved; thus $A\cdot B = A_xB_x + A_yB_y + A_zB_z$, $(A\times B)_x = A_yB_z - A_zB_y$, etc.

[2] The terms in these formulas are not operators, but ordinary vectors and scalars; in none of the following expressions is $\nabla$ to be considered as operating on an unwritten argument as hitherto. Each term in the following formulas should properly be enclosed in braces.

* NATIONAL RESEARCH FELLOWS.

# PROJECTIVE DIFFERENTIATION OF SPINORS

BY OSWALD VEBLEN AND A. H. TAUB

THE INSTITUTE FOR ADVANCED STUDY AND PRINCETON UNIVERSITY

Communicated December 13, 1933

1. In this note we develop the theory of covariant differentiation[1] of the spinors which were recently defined, introduce the assumptions about the covariant derivatives of the fundamental spinors of the projective relativity theory which are needed for a smoothly working calculus, and establish the consistency[2] of these assumptions.

The definition of covariant differentiation is an obvious generalization of that used in tensor analysis. It is based upon an arbitrary *spin connection* whose components are complex functions of the coördinates $(x^1, \ldots x^4)$ obeying the law of transformation,

$$\bar{\Gamma}^{A}_{B\alpha} = \left(\Gamma^{C}_{D\beta}\, t^{D}_{B} + \frac{\partial t^{C}_{B}}{\partial x^{\beta}}\right) T^{A}_{C} \frac{\partial x^{\beta}}{\partial \bar{x}^{\alpha}}, \tag{1.1}$$

and an arbitrary projective connection which has the law of transformation

$$\bar{\Gamma}^{\alpha}_{\beta\gamma} = \left(\Gamma^{\lambda}_{\mu\nu} \frac{\partial x^{\mu}}{\partial \bar{x}^{\beta}} \frac{\partial x^{\nu}}{\partial \bar{x}^{\gamma}} + \frac{\partial^2 x^{\lambda}}{\partial \bar{x}^{\beta} \partial \bar{x}^{\gamma}}\right) \frac{\partial \bar{x}^{\alpha}}{\partial x^{\lambda}}. \tag{1.2}$$

Typical formulae of covariant differentiation then are

$$\psi^{A}_{,\alpha} = \frac{\partial \psi^{A}}{\partial x^{\alpha}} + \Gamma^{A}_{B\alpha}\, \psi^{B} - \frac{1}{4}\, \Gamma^{B}_{B\alpha}\, \psi^{A}, \tag{1.3}$$

where $\psi^A$ is a spinor of weight $1/4$,

$$\gamma_{AB,\,\alpha} = \frac{\partial \gamma_{AB}}{\partial x^{\alpha}} - \gamma_{AC}\Gamma^{C}_{B\alpha} - \gamma_{CB}\Gamma^{C}_{A\alpha} + \frac{1}{2}\, \Gamma^{C}_{C\alpha}\gamma_{AB} \tag{1.4}$$

where $\gamma_{AB}$ is of weight $-1/2$,

$$\gamma^{\alpha}_{;\lambda} = \frac{\partial \gamma^{\alpha}}{\partial x^{\lambda}} + \Gamma_{\lambda}\gamma^{\alpha} - \gamma^{\alpha}\Gamma_{\lambda} + \Gamma^{\alpha}_{\mu\lambda}\gamma^{\mu}, \tag{1.5}$$

where $\gamma^{\alpha}$ is the matrix notation for the spinor $\gamma^{A\alpha}_{B}$ of weight zero and where we have used matrix multiplication. The usual elimination process shows that the covariant derivatives thus defined are spinors.

The formulae for the covariant derivatives of spinors with dotted indices can be obtained by the same procedure. Thus

$$\varphi^{\dot{B}}_{;\lambda} = \frac{\partial \varphi^{\dot{B}}}{\partial x^{\lambda}} + \varphi^{\dot{C}}\Lambda^{\dot{B}}_{\dot{C}\lambda} - \frac{1}{4}\, \Lambda^{\dot{C}}_{\dot{C}\lambda}\varphi^{\dot{B}}$$

where $\varphi^{\dot{B}}$ is a spinor of conjugate weight $^1/_4$ and the quantities $\Lambda^{\dot{B}}_{C\lambda}$ transform like the complex conjugates of $\Gamma^{B}_{C\lambda}$. Since the laws of transformation of $\varphi^{\dot{B}}$ and $\psi^{B*}$ are the same if their weights are the same, we will have

$$\psi^{B*}_{,\lambda} = \frac{\partial \psi^{B*}}{\partial x^{\lambda}} + \psi^{C*}\Lambda^{\dot{B}}_{\ C\lambda} - \frac{1}{4}\Lambda^{\dot{C}}_{\ \dot{C}\lambda}\,\psi^{B*}.$$

We now make the requirement that the operation of taking the complex conjugate be commutative with the operation of covariant differentiation, that is, $(\psi^{B}_{,\lambda})^* = (\psi^{B*})_{,\lambda}$. It follows that $\Lambda^{\dot{B}}_{C\lambda} = \Gamma^{B*}_{C\lambda}$. Under this assumption we have as a typical example,

$$\gamma_{,\alpha} = \frac{\partial \gamma}{\partial x^{\alpha}} + \Gamma_{\alpha}\gamma - \gamma\Gamma_{\alpha}^{*} - \frac{1}{4}\Gamma^{C}_{C\alpha}\gamma + \frac{1}{4}\Gamma^{C*}_{C\alpha}\,\gamma, \qquad (1.6)$$

where $\gamma$ is the matrix notation for the fundamental spinor $\gamma^{A}_{\dot{B}}$ of direct weight $+\,^1/_4$ and conjugate weight $-\,^1/_4$.

2. Other commutation requirements will impose further conditions on both the spin and projective connections. Thus, if we require that the process of changing from dotted to undotted indices be commutative with covariant differentiation, i.e., if relations of the type $\psi^{\dot{A}}_{,\lambda} = \gamma^{\dot{A}}_{B}\,\psi^{B}_{,\lambda}$ are to be satisfied, we must have

$$\gamma_{,\gamma} = 0. \qquad (2.1)$$

Similarly the requirement that covariant differentiation be interchangeable with the process of raising and lowering indices, i.e., that relations of the type $\varphi_{\alpha,\lambda} = \gamma_{\alpha\beta}\,\varphi^{\beta}_{,\lambda}$ and $\psi_{A,\lambda} = \gamma_{AB}\,\psi^{\beta}_{,\alpha}$ are to be satisfied, implies that

$$\gamma_{\alpha\beta,\lambda} = 0 \qquad (2.2) \qquad \text{and} \qquad \gamma_{AB,\lambda} = 0. \qquad (2.3)$$

Finally, if covariant differentiation is to be commutative with the process of shifting from spin indices to projective indices, i.e., if relations of the type $X^{\alpha}_{;\beta} = \gamma^{A\alpha}_{B}\,X^{B}_{A,\beta}$ are to be satisfied, we must have

$$\gamma^{\alpha}_{;\lambda} = 0 \qquad (2.4)$$

3. Equations (2.2) and (2.3) are consequences of (2.4). From the relation

$$\frac{1}{2}(\gamma^{\alpha}\gamma^{\beta} + \gamma^{\beta}\gamma^{\alpha}) = \frac{1}{4}\gamma^{\alpha\beta}.1 \qquad (3.1)$$

it is immediately evident that (2.2) holds if (2.4) does. That (2.3) is a consequence of (2.4) can be seen from the relation (10.6) p. 987, Vol. 19 (1933) of these PROCEEDINGS.

$$\gamma^{A}_{B\alpha}\gamma^{E\alpha}_{F} = -\frac{1}{4}\,\delta^{A}_{B}\delta^{E}_{F} + \frac{1}{2}\,\delta^{E}_{B}\delta^{A}_{F} + \frac{1}{2}\,\gamma^{EA}\gamma_{BF}. \tag{3.2}$$

Taking the covariant derivative of both sides we obtain

$$\gamma^{EA}_{,\lambda}\gamma_{BF} + \gamma^{EA}\gamma_{BF,\,\lambda} = 0.$$

Multiplying by $\gamma_{EA}$ and summing, we obtain $4\gamma_{BF,\lambda} = \gamma_{EA}\,\gamma^{EA}_{,\lambda}\gamma_{BF}$. However,

$$\gamma^{EA}_{,\lambda} = \frac{\partial\gamma^{EA}}{\partial x^{\lambda}} + \gamma^{CA}K^{E}_{C\lambda} + \gamma^{EC}\,K^{A}_{C\lambda} \tag{3.3}$$

where $K^{E}_{C\lambda} = \Gamma^{E}_{C\lambda} - \frac{1}{4}\,\delta^{E}_{C}\,\Gamma^{F}_{F\lambda}$ is the traceless part[3] of $\Gamma_{\lambda}$. Hence

$$\gamma_{EA}\gamma^{EA}_{,\lambda} = \gamma_{EA}\,\frac{\partial\gamma^{EA}}{\partial x^{\lambda}} - 2K^{C}_{C\lambda}.$$

Since both terms on the right-hand side of this equation are zero, it follows that $\gamma_{BF,\,\lambda} = 0$. Another consequence of (2.4) is obtained from the relation

$$\frac{1}{2}\,(\gamma^{\alpha}\gamma^{\beta} - \gamma^{\beta}\gamma^{\alpha}) = s^{\alpha\beta}. \tag{3.4}$$

It is

$$s^{\alpha\beta}_{,\lambda} = 0. \tag{3.5}$$

4. Equation (2.4) enables us to obtain relations between the spin and projective connections. Writing equation (2.4) in full we have

$$\frac{\partial\gamma^{\alpha}}{\partial x^{\lambda}} + K_{\lambda}\gamma^{\alpha} - \gamma^{\alpha}K_{\lambda} + \gamma^{\mu}\Gamma^{\alpha}_{\mu\lambda} = 0. \tag{4.1}$$

Multiplying by $\gamma_{\beta}$ and taking the trace, this becomes

$$\Gamma^{\alpha}_{\beta\lambda} = -\,\text{Trace}\left[2s^{\alpha}_{\beta}\,K_{\lambda} + \gamma_{\beta}\,\frac{\partial\gamma^{\alpha}}{\partial x^{\lambda}}\right]$$

$$= -\,\text{Trace}\left[2s^{\alpha}_{\beta}\,\Gamma_{\lambda} + \gamma_{\beta}\frac{\partial\gamma^{\alpha}}{\partial x^{\lambda}}\right] \tag{4.2}$$

Using the relation

$$-\,2s^{A\alpha}_{B\beta}\gamma^{E}_{F\alpha}\gamma^{F\beta}_{G} = \frac{1}{2}\,\delta^{A}_{G}\delta^{E}_{B} + \gamma^{AE}\gamma_{BG}$$

we can obtain from (4.2) an expression giving the spin connection in terms of the projective connection. Thus multiplying equation (4.2) by $\gamma^{E}_{F\alpha}\gamma^{F\beta}_{G}$ and summing, we have

$$\frac{1}{2}(\delta^A_G\delta^E_B + \gamma^{AE}\gamma_{BG})\ K^B_{A\lambda} = \Gamma^\alpha_{\beta\lambda}\gamma^E_{F\alpha}\ \gamma^{F\beta}_G + \gamma^E_{F\alpha}\frac{\partial\gamma^{F\alpha}_G}{\partial x^\lambda} + \gamma^{BE}\frac{\partial\gamma_{BG}}{\partial x^\lambda}. \quad (4.3)$$

Since $\gamma^{AB}_{,\lambda} = 0$ is a consequence of (4.1) we may use equation (3.3) to evaluate the second term on the left-hand side. From it we obtain

$$\gamma^{AE}\gamma_{BG}\ K^B_{A\lambda} = K^E_{G\lambda} + \gamma_{BG}\frac{\partial\gamma^{EB}}{\partial x^\lambda}.$$

Hence we have

$$K_\lambda = \Gamma^\alpha_{\beta\lambda}\gamma_\alpha\gamma^\beta + \gamma_\alpha\frac{\partial\gamma^\alpha}{\partial x^\lambda} + \frac{1}{2}\eta_\lambda \quad (4.4)$$

where $\gamma^{BE}\partial\gamma_{BG}/\partial x^\lambda = \eta^E_{G\lambda}$.

Another form of this equation is obtained when we use equation (4.1) to eliminate the term $\gamma^{AE}\gamma_{BG}\ K^B_{A\lambda}$ from equation (4.3). To do this we multiply on the right by $\gamma_\alpha$ and use equation (3.2). We then have

$$\frac{1}{2}(3\delta^E_B\delta^A_G - \gamma^{AE}\gamma_{BG})\ K^B_{A\lambda} = -\Gamma^\alpha_{\beta\lambda}\gamma^{E\beta}_F\gamma^F_{G\alpha} - \gamma^F_{G\alpha}\frac{\partial\gamma^{E\alpha}_F}{\partial x^\lambda}. \quad (4.5)$$

Adding equations (4.3) and (4.5) we obtain

$$K_\lambda = -\Gamma^\alpha_{\beta\lambda}\ s^\beta_\alpha + \gamma_\alpha\frac{\partial\gamma^\alpha}{\partial x^\lambda} + \frac{1}{2}\eta_\lambda. \quad (4.6)$$

5. The condition (2.2) is known to be equivalent to the statement that $\Gamma^\alpha_{\beta\lambda}$ is expressible as follows:

$$\Gamma^\alpha_{\beta\lambda} = \left\{{\alpha \atop \beta\lambda}\right\} + \gamma^{\rho\alpha}A_{\beta\lambda\rho} = \left\{{\alpha \atop \beta\lambda}\right\} + A^\alpha_{\beta\lambda} \quad (5.1)$$

where $A^\alpha_{\beta\lambda}$ is a projective tensor such that $A_{\beta\lambda\rho} = -A_{\rho\lambda\beta}$ and $\left\{{\alpha \atop \beta\lambda}\right\}$ are the Christoffel symbols of the second kind formed from the $\gamma_{\alpha\beta}$.

Equation (2.2) leaves the $A_{\beta\lambda\rho}$ completely arbitrary. We shall now show that they are still arbitrary when we assume (2.4). We substitute equations (4.6) into

$$\gamma^\mu_{,\lambda} = \frac{\partial\gamma^\mu}{\partial x^\lambda} + K_\lambda\gamma^\mu - \gamma^\mu K_\lambda + \gamma^\beta\Gamma^\mu_{\beta\lambda} \quad (5.2)$$

and obtain

$$\gamma^\mu_{,\lambda} = \frac{\partial\gamma^\mu}{\partial x^\lambda} + \left[-\Gamma^\beta_{\alpha\lambda}\ s^\alpha_\beta + \gamma_\alpha\frac{\partial\gamma^\alpha}{\partial x^\lambda} + \frac{1}{2}\eta_\lambda\right]\gamma^\mu$$

$$- \gamma^{\mu}\left[ - \Gamma^{\beta}_{\alpha\lambda} s^{\alpha}_{\beta} + \gamma_{\alpha} \frac{\partial \gamma^{\alpha}}{\partial x^{\lambda}} + \frac{1}{2} \eta_{\lambda} \right] + \gamma^{\beta} \Gamma^{\mu}_{\beta\lambda}$$

$$= \frac{1}{2} \gamma_{\rho} \frac{\partial \gamma^{\mu\rho}}{\partial x^{\lambda}} + [\gamma^{\mu} s^{\alpha}_{\beta} - s^{\alpha}_{\beta} \gamma^{\mu} + \delta^{\mu}_{\beta} \gamma^{\alpha}] \Gamma^{\beta}_{\alpha\lambda}. \quad (5.3)$$

If we now make use of the relation

$$[\gamma^{\mu} s^{\alpha}_{\beta} - s^{\alpha}_{\beta} \gamma^{\mu} + \delta^{\mu}_{\beta} \gamma^{\alpha}] = \frac{1}{2} (\gamma_{\beta} \gamma^{\alpha\mu} + \gamma^{\alpha} \delta^{\mu}_{\beta}) \quad (5.4)$$

equation (5.3) becomes

$$\gamma^{\mu}_{,\lambda} = \frac{1}{2} \gamma_{\rho} \left( \frac{\partial \gamma^{\mu\rho}}{\partial x^{\lambda}} + \gamma^{\alpha\mu} \Gamma^{\rho}_{\alpha\lambda} + \gamma^{\rho\alpha} \Gamma^{\mu}_{\alpha\lambda} \right) = \frac{1}{2} \gamma_{\rho} \gamma^{\mu\rho}_{,\lambda}.$$

Since $\gamma^{\mu\rho}_{,\lambda} = 0$ implies $\gamma_{\mu\rho,\lambda} = 0$ and conversely, we have that equations (5.1) and (4.6) imply equation (2.4).

However (5.1) leaves the $A$'s completely arbitrary and (4.6) determines the spin connection if the projective connection is given. Hence we have shown that equations (2.4) can be satisfied when the $A$'s are arbitrary, provided the spin connection is determined by (4.6). Since (4.6) may be derived from (2.4), we see that (2.4) is consistent with an arbitrary choice of the $A$'s. Also since (2.2) and (2.3) are derivable from (2.4) they are consistent with an arbitrary choice of the $A$'s.

We have also shown that equations (5.1) and (4.6) are necessary and sufficient conditions for equation (2.4), i.e., equations (5.1) and (4.6) are fully equivalent to (2.4).

6. Another set of necessary and sufficient conditions for (4.1) are obtained by substituting equations (4.2) into (4.1). Thus

$$\gamma^{\alpha}_{,\lambda} = \frac{\partial \gamma^{\alpha}}{\partial x^{\lambda}} + K_{\lambda} \gamma^{\alpha} - \gamma^{\alpha} K_{\lambda} - \gamma^{\beta} \operatorname{Trace} \left( 2K_{\lambda} s^{\alpha}_{\beta} + \gamma_{\beta} \frac{\partial \gamma^{\alpha}}{\partial x^{\lambda}} \right). \quad (6.1)$$

Since

$$-2\gamma^{A\beta}_{B} s^{D\alpha}_{C\beta} = \frac{1}{2} \gamma^{E\alpha}_{C} (\delta^{A}_{E} \delta^{D}_{B} + \gamma^{AD} \gamma_{EB}) - \frac{1}{2} \gamma^{D\alpha}_{E} (\delta^{A}_{C} \delta^{E}_{B} + \gamma^{AE} \gamma_{BC}) \quad (6.2)$$

(6.1) becomes

$$\gamma^{A\alpha}_{B,\lambda} = \frac{1}{2} \left[ \gamma^{C\alpha}_{D} \frac{\partial (\gamma^{AD} \gamma_{CB})}{\partial x^{\lambda}} - \gamma^{A\alpha}_{D} K^{D}_{B\lambda} - \gamma^{D\alpha}_{E} \gamma^{AE} \gamma_{BC} K^{C}_{D\lambda} \right.$$

$$+ \gamma_C^{E\alpha}\gamma^{AD}\gamma_{EB}\,K_{D\lambda}^{C} + \gamma_B^{D\alpha}K_{D\lambda}^{A}\Big]$$

$$= \frac{1}{2}\,\gamma_D^{C\alpha}(\gamma^{AD}\gamma_{BC}),_{\lambda}$$

Hence equations (4.2) and (2.3) are necessary and sufficient conditions for (2.4). Therefore (4.2) and (2.3) are fully equivalent to (2.4).

7. We are now ready to discuss the reality of the projective connection $\Gamma_{\beta\lambda}^{\alpha}$. From equations (4.2) we have

$$\Gamma_{\beta\lambda}^{\alpha *} - \Gamma_{\beta\lambda}^{\alpha} = -2\left(K_{B\lambda}^{A*}\gamma_C^{\dot{B}}\gamma_{\dot{A}}^{D} - K_{C\lambda}^{D} + \frac{\partial\gamma_C^{\dot{A}}}{\partial x^{\lambda}}\gamma_{\dot{A}}^{D}\right)s_{D\beta}^{C\alpha}$$

$$= -2\gamma_{\dot{A}}^{D}\gamma_{C,\lambda}^{\dot{A}}\,s_{D\beta}^{C\alpha} \qquad (7.1)$$

where we have used the relations $s_{B\beta}^{A\alpha *} = s_{B\beta}^{\dot{A}\alpha}$ and $\gamma_B^{\dot{A}\alpha} = \gamma_B^{A\alpha *}$. From (7.1) we see that (2.1) is a sufficient condition for the reality of $\Gamma_{\beta\lambda}^{\alpha}$.

It is not a necessary condition. From equation (4.6) we have

$$K_{D\lambda}^{A*} = -\Gamma_{\beta\lambda}^{\alpha}s_{\dot{D}\alpha}^{\dot{A}\beta} + \gamma_{\dot{F}\alpha}^{\dot{A}}\frac{\partial\gamma_{\dot{D}}^{\dot{F}\alpha}}{\partial x^{\lambda}} + \frac{1}{2}\,\gamma^{\dot{F}A}\frac{\partial\gamma_{\dot{F}\dot{D}}}{\partial x^{\lambda}}.$$

Now

$$\gamma_{\dot{F}\alpha}^{\dot{A}}\frac{\partial\gamma_{\dot{D}}^{\dot{F}\alpha}}{\partial x^{\lambda}} = \gamma_C^{\dot{A}}\gamma_{\dot{D}}^{B}\gamma_{E\alpha}^{C}\frac{\partial\gamma_B^{E\alpha}}{\partial x^{\lambda}} + \frac{3}{2}\,\gamma_B^{\dot{A}}\frac{\partial\gamma_{\dot{D}}^{B}}{\partial x^{\lambda}} + \frac{1}{2}\,\gamma_{\dot{F}\dot{D}}\,\gamma^{L\dot{A}}\frac{\partial\gamma_L^{\dot{F}}}{\partial x^{\lambda}} + \frac{1}{2}\,\delta_{\dot{D}}^{\dot{A}}\gamma_{\dot{F}}^{L}\frac{\partial\gamma_L^{\dot{F}}}{\partial x^{\lambda}}$$

and

$$\frac{1}{2}\,\gamma^{\dot{F}\dot{A}}\frac{\partial\gamma_{\dot{F}\dot{D}}}{\partial x^{\lambda}} = \frac{1}{2}\,\gamma_C^{\dot{A}}\gamma_{\dot{D}}^{B}\gamma^{LC}\frac{\partial\gamma_{LB}}{\partial x^{\lambda}} - \frac{1}{2}\,\gamma_B^{\dot{A}}\frac{\partial\gamma_{\dot{D}}^{B}}{\partial x^{\lambda}} - \frac{1}{2}\,\gamma_{\dot{F}\dot{D}}\,\gamma^{L\dot{A}}\frac{\partial\gamma_L^{\dot{F}}}{\partial x^{\lambda}}.$$

Hence

$$K_{D\lambda}^{A*} = K_{B\lambda}^{C}\gamma_C^{\dot{A}}\gamma_{\dot{D}}^{B} + \gamma_B^{\dot{A}}\frac{\partial\gamma_{\dot{D}}^{B}}{\partial x^{\lambda}} + \frac{1}{2}\,\delta_{\dot{D}}^{\dot{A}}\gamma_{\dot{F}}^{L}\frac{\partial\gamma_L^{\dot{F}}}{\partial x^{\lambda}}.$$

Multiplying by $\gamma_{\dot{A}}^{F}$ and summing, we have

$$\gamma,_{\lambda} = \frac{\partial\gamma}{\partial x^{\lambda}} + K_{\lambda}\,\gamma - \gamma K_{\lambda}^{*} = \gamma f_{\lambda}$$

where $f_{\lambda} = \text{Trace}\left(\gamma\dfrac{\partial\gamma}{\partial x^{\lambda}}\right)$. Hence $\gamma,_{\lambda} = \gamma f_{\lambda}$ is a necessary condition for the reality of $\Gamma_{\beta\lambda}^{\alpha}$. By (7.1) it is sufficient.

Hence the conditions (2.1), (2.2), (2.3) and (2.4) are self consistent and imply that $A_{\beta\lambda\mu}$ is real and anti-symmetric in $\beta$ and $\mu$ but otherwise arbitrary.

[1] This theory is the counterpart of that to be found in varying degrees of maturity in the papers cited in the notes to "Spinors in Projective Relativity," these PROCEEDINGS, **19**, pp. 979–989 (1933). To these references should have been added V. Bargmann. Bemerkungen zur allgemein-relativistischen Fassung der Quantentheorie, *Berl. Ber*, **1932**, pp. 346–354, in which there is introduced and discussed a matrix, $\alpha$, somewhat analogous to $\gamma$. Also there should now be added reference to two papers on projective relativity and spinors which have just appeared: W. Pauli, Ueber die Formulierung der Naturgesetze mit fuenf homogenen Koordinaten, Parts I and II, *Ann. der Phys.*, **18** (5 Folge, 1933), pp. 305–336 and pp. 337–372. Pauli seems to attach importance to the use of homogeneous coördinates $x^0, x^1, \ldots, x^4$ (Van Dantzig) to represent the underlying four-dimensional space-time of projective relativity. But these coördinates are derived from the coördinates $x^1, \ldots, x^4$ and the gauge variable $x^0$ used by Veblen and Hoffmann, *Phys. Rev.*, **36**, p. 810 (1930) by the transformation

$$X^0 = e^{x^0}$$
$$X^i = e^{x^0} x^i. \qquad (i = 1, \ldots, 4)$$

Such a transformation can make no difference either in the physics or the geometry of the theory. It seems to us to make things appear more complicated.

In a five-dimensional representation the use of the homogeneous coördinates $(X^0, \ldots, X^5)$ amounts to representing the points of space-time by the straight lines through the origin, whereas the use of $x^1, \ldots, x^4$, and the gauge variable amounts to using the system of straight lines parallel to the $x^0$-axis for the same purpose. The transformation given above carries the one system of lines into the other.

There is an analogous relation between the Einstein-Mayer unified theory of 1931 and the projective relativity. Whereas the projective relativity used homogeneous coördinates in the tangent 4-spaces, Einstein and Mayer accomplish essentially the same thing by imaging each tangent 4-space upon a system of parallel straight lines in a 5-space. (Cf. Chap. VIII, Projektive Relativitätstheorie, by O. Veblen, Berlin, 1933.) Our preference for ordinary non-homogeneous coördinates in space-time and homogeneous coördinates in the tangent spaces corresponds to the fact that one always deals with a limited region of space-time, whereas the projective tangent space has to be used in its entirety.

[2] In this we have been greatly helped by questions and remarks by Professor von Neumann in a seminar in which this problem has been discussed.

[3] The occurrence of $K_\alpha$ in the subsequent work instead of $\Gamma_\alpha$ is due to the manner in which the weights were assigned to the fundamental spinors, $\gamma^\alpha$, $\gamma$ and $\gamma^{AB}$. The quantities $K_\alpha$ may be called the components of a spin connection of the second kind (cf. these PROCEEDINGS, **19**, p. 472 (1933)). The law of transformation of the $K_\alpha$'s is

$$\bar{K}^A_{B\lambda} = \left( K^C_{D\beta}\, t^D_B + \frac{\partial t^C_B}{\partial x^\beta} \right) T^A_C \frac{\partial x^\beta}{\partial \bar{x}^\lambda} - \frac{1}{4} \frac{\partial \log t}{\partial x^\beta} \frac{\partial x^\beta}{\partial \bar{x}^\alpha}.$$

The last term arises from the law of transformation of $\Gamma^c_{c\lambda}$, which is

$$\bar{\Gamma}^c_{c\lambda} = \left( \bar{\Gamma}^c_{c\beta} + \frac{\partial \log t}{\partial x^\beta} \right) \frac{\partial x^\beta}{\partial \bar{x}^\lambda}.$$

This is the law of transformation of a projective vector, with the invariant condition $\Gamma_0 = 1$, under the transformation

$$\begin{aligned} \bar{x}^i &= \bar{x}^i(x^1 \ldots x^4) \\ \bar{x}^0 &= x^0 + \log \rho(x^1 \ldots x^4) \end{aligned} \qquad i = 1, \ldots, 4$$

if we assume that $t = \rho$. This shows that the spinor theory can be specialized so that the gauge transformations are connected with the spin transformations by the relation $\rho = t$.